Process optimization of friction stir welding based on thermal models

Larsen, Anders Astrup

Publication date:
2010

Document Version
Early version, also known as pre-print

Process optimization of friction stir welding based on thermal models

by

Anders Astrup Larsen

Department of Mathematics

Technical University of Denmark

Title of Thesis:
Process optimization of friction stir welding based on thermal models

Ph.D. student:
Anders Astrup Larsen
Department of Mathematics
Technical University of Denmark
Address: Matematiktorvet, DTU Building 303 S, DK-2800 Lyngby, Denmark
E-mail: a.a.larsen@mat.dtu.dk

Supervisors:
Mathias Stolpe
Department of Mathematics
Technical University of Denmark
Address: Matematiktorvet, DTU Building 303 S, DK-2800 Lyngby, Denmark
E-mail: m.stolpe@mat.dtu.dk

Jesper Hattel
Department of Mechanical Engineering
Technical University of Denmark
Address: Nils Koppels Allé, DTU Building 425, DK-2800 Lyngby, Denmark
E-mail: jhat@mek.dtu.dk

Ole Sigmund
Department of Mechanical Engineering, Solid Mechanics
Technical University of Denmark
Address: Nils Koppels Allé, DTU Building 404, DK-2800 Lyngby, Denmark
E-mail: sigmund@mek.dtu.dk

Preface

This thesis is submitted in partial fulfillment of the requirements for obtaining the degree of Ph.D. at the Technical University of Denmark (DTU). The Ph.D. project was funded partly by the Danish Research Council for Technology and Production Sciences and partly by DTU and carried out as a collaboration between the Department of Mathematics and the Department of Mechanical Engineering at DTU in the period June 15^{th} 2006 - June 14^{th} 2009. Supervisors on the project were associate professor Mathias Stolpe from the Department of Mathematics, professor Jesper Hattel and professor dr. techn. Ole Sigmund from the Department of Mechanical Engineering.

I wish to thank my supervisors and Martin P. Bendsøe, who was my main supervisor for the first two and a half years, for an interesting project and many fruitful meetings and discussions during the past three years. Furthermore, I thank associate professor Henrik Schmidt from the Department of Mechanical Engineering for the many meetings and much advice during the three years. I also thank professor Robert B. Haber, his ph.d. students Scott Miller and Reza Abedi from University of Illinois at Urbana-Champaign in USA, as well as assistent professor Domenico Lahaye from the Technical University of Delft in The Netherlands for good collaboration and for making my visits there possible.

Finally, I thank my colleagues at the Department of Mathematics and the TopOpt Group for a very positive working environment during the past three years.

Kgs. Lyngby, June 2009

Anders Astrup Larsen

Resumé (in Danish)
Proces optimering af friction stir svejsning baseret på termiske modeller

Denne afhandling undersøger hvordan optimeringsmetoder kan anvendes på numeriske modeller af en friction stir svejseproces. Intentionen bag dette arbejde er at vise anvendeligheden af forskellige metoder, som kan bruges i forbindelse med komplekse, muligvis tidskrævende, modeller hvor effektiv beregning af gradienter ikke er mulig. Fokus er derfor på surrogat optimeringsmetoder med det formål at reducere antallet af evalueringer af den dyre model, ved brug af en mindre præcis model sammen med den præcise model, som søges optimeret. De anvendte metoder kræver ikke, at brugeren bestemmer gradienter af den præcise model.

Optimeringsmetoderne er anvendt på stationære termiske modeller, af varierende kompleksitet, af friction stir svejseprocessen. De undersøgte optimeringsproblemer er baseret på optimering af temperaturfeltet i svejseemnet ved bestemmelse af optimal svejse- og rotationshastighed. Udover det deterministiske tilfælde, er også et robust optimeringsproblem, hvor der tages hensyn til effekten af usikkerheder i materiale- og optimeringsparametre, undersøgt. Målet er at opnå et givet middelrespons samtidig med at standard afvigelsen af responset reduceres. Desuden er et optimeringsproblem baseret på en mikrostruktur model, som muliggør beregning af hårdheden af den svejsede plade, løst. Brugen af rent termiske modeller er en forenkling i forhold til den virkelige proces, men alligevel viser det anvendeligheden af optimeringsmetoderne og danner basis for optimering af mere detaljerede modeller. Surrogat modeller af varierende kompleksitet, og dermed lighed med den rigtige model, er anvendt og betydningen for resultaterne er diskuteret.

Desuden bidrager denne afhandling til modelleringen af varmetabet mellem svejseemnet og den underliggende plade ved løsning af et inverst problem, hvori eksperimentelle data og en numerisk model bruges til bestemmelse af varmeovergangstallet. Forskellige parametriseringer af fordelingen af varmeovergangstallet er undersøgt og optimeringsproblemet er formuleret som en minimering af forskellen mellem målte og beregnede temperaturer. Størrelsen og fordelingen af varmeovergangstallet er bestemt for de tilgængelige data.

Summary
Process optimization of friction stir welding based on thermal models

This thesis investigates how to apply optimization methods to numerical models of a friction stir welding process. The work is intended as a proof-of-concept using different methods that are applicable to models of high complexity, possibly with high computational cost, and without the possibility for efficient gradient calculation. Thus, the focus is on surrogate optimization methods with the aim of reducing the number of expensive function evaluations, by using a low-fidelity model together with the high-fidelity model to be optimized. The methods used here do not require the user to supply gradient information of the high-fidelity model.

The optimization schemes are applied to stationary thermal models of differing complexity of the friction stir welding process. The optimization problems considered are based on optimizing the temperature field in the workpiece by finding optimal translational speed and rotational speed of the tool. Besides the deterministic problem a robust optimization problem is considered in which the effects of uncertain material and optimization parameters are taken into account. The objective is to obtain a desired mean response while reducing the standard deviation of the response. Also an optimization problem based on a microstructure model is solved, allowing the hardness distribution in the plate to be optimized. The use of purely thermal models represents a simplification of the real process; nonetheless, it shows the applicability of the optimization methods considered and forms the basis for optimization of more detailed models. Surrogate models of varying complexity, and similarity with the true model, are applied and the effect on the optimization results is discussed.

Furthermore, the thesis contributes to the modelling of the heat transfer between the workpiece and the backingplate by solving an inverse modelling problem in which experimental data and a numerical model are used for determining the contact heat transfer coefficient. Different parametrizations of the spatial distribution of the heat transfer coefficient are studied and discussed, and the optimization problem is formulated as a minimization of the difference between measured and calculated temperatures. The magnitude and distribution of the heat transfer coefficient is determined for the available data.

Contents

Preface **i**

Resumé (in Danish) **ii**

Summary **iii**

Contents **iv**

1 Modelling of friction stir welding **2**
1.1 Thermal models 4
1.1.1 Heat source models 4
1.1.2 Analytical models 8
1.1.3 Numerical models 11
1.2 Mechanical and material aspects of FSW 14
1.2.1 Residual stresses and flow models 14
1.2.2 Material hardness 15

2 Optimization techniques **18**
2.1 Surrogate techniques 20
2.1.1 Additive and multiplicative correction 20
2.1.2 Space mapping techniques 21
2.1.3 Manifold mapping techniques 25
2.1.4 Response surface techniques 29
2.2 Robust optimization 33

3 Summary of results **38**
3.1 Contributions and impact 39

4 Concluding remarks **41**
4.1 Future work 42

References **43**

A Article 1 **49**

B Article 2 **61**

C Article 3 **76**

D Article 4 **86**

E Maximization of advance per revolution **99**

F Trust-region and filter methods **101**

Introduction

The present thesis deals with numerical optimization of process parameters of a friction stir welding (FSW) process. This means applying optimization techniques to models of the process in order to optimize certain aspects of the process by finding optimal model parameters. Thus we do not deal with actual physical experiments, although some of the optimization techniques considered, e.g. response surface methods, could be used if the mathematical model was replaced by an experiment. The work is intended as a proof-of-concept by studying different optimization techniques and their applicability to optimization of FSW.

FSW is a coupled thermo-mechanical process leading to very complex models if all aspects, such as heat transfer, contact conditions, etc., are taken into account. In many cases however it is reasonable to model only parts of the physics and use for example purely thermal models. Chapter 1 describes the friction stir welding process and different types of models, with emphasis on thermal models as they form the basis for the optimization problems considered in this work.

Generally, optimization is an iterative process that requires repeated analyses of the underlying model that is to be optimized. This means that for time consuming models, as is often the case in process modelling, it is desirable to use techniques that require few evaluations of the model. This is the main motivation for using the surrogate modelling techniques presented in Chapter 2.

Chapter 3 briefly summarizes the results presented in the articles, and the contributions of the present thesis, while Chapter 4 concludes the thesis and discusses directions for future work. The appendices contain the four articles included, an example of a friction stir welding optimization problem and a note on trust region and filter methods used in connection to the optimization techniques in Chapter 2.

The numerical models presented in the thesis are all solved using the finite element programme Comsol Multiphysics, Comsol (2006), in combination with Matlab, Matlab (2006). The optimization problems are solved using, in most cases, SNOPT, Gill et al. (2006), or, in a few cases, the optimizers available in Matlab.

Chapter 1

Modelling of friction stir welding

In this chapter a number of models of the friction stir welding (FSW) process are presented as they form the basis for the optimization problems considered in this work. The main focus is on thermal models while other types of models are briefly discussed.

FSW is a solid state welding process invented in 1991 by The Welding Institute, TWI (2007). Saying that the process is solid state means that there is no, or very limited and localised, melting of the material as opposed to conventional methods like arc welding where a central aspect is the melting of the material. The process is affected by many parameters, such as the material properties of the workpiece, e.g., the thermal conductivity and heat capacity, the environment, for example the temperature of the surroundings, the tool design and other factors. Once the welding setup, such as workpiece dimensions and material, has been set, the main process parameters to be chosen, in order to obtain a successful weld, are the tool design and the translational speed u_w and the rotational speed ω. The present work deals primarily with the use of optimization techniques for optimal selection of u_w and ω. In this chapter more details on the FSW process and mathematical modelling of FSW are presented.

The process is sketched in Figure 1.1 which shows the welding setup, including the tool, the workpiece and the backingplate, and some terms commonly used in connection to FSW. The tool consists of a cylinder, that is in contact with the top surface of the workpieces, and a probe (sometimes called the pin) of smaller diameter that is forced, by an axial load, into the material between the plates to be welded. The contact area between the cylinder and the workpieces is denoted the shoulder and this contact zone is responsible for a large part of the total heat generation. The probe contributes less to the total heat generation but has the effect of stirring the material from the two plates to create the weld. Many different tool designs exist, with the simplest one being a cylindrical probe attached to a flat shoulder. More complex tools may have conical shoulders and threaded or triangular probe designs, see for example Colegrove et al. (2003) where different tool shapes are compared using experiments and computational fluid dynamics models.

The welding process may be divided into four phases. First the *plunge* action where the rotating, but otherwise stationary, tool is forced into the weldline in the workpiece. When the shoulder of the tool contacts the workpiece surface the tool is kept stationary for a short time known as the *dwell* period where the workpiece gradually heats up

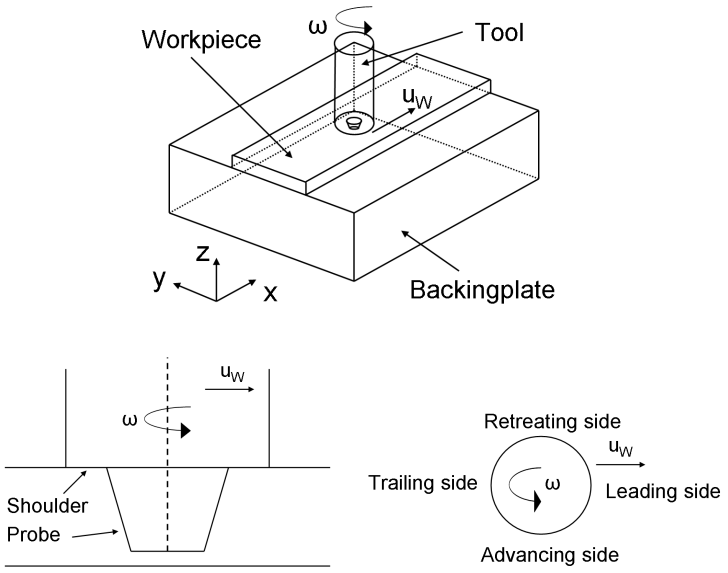

Figure 1.1: Top: The friction stir welding setup consisting of the workpiece, the tool and the backingplate. Bottom left: Cross section of the plate with the tool. The figure shows a tool with a flat shoulder and a conical probe, but many other tool designs exist. Bottom right: Terms denoting positions relative to the tool.

and the material surrounding the tool is softened. After this, the tool is traversed along the weld line (*welding* period) to join the two parts before the transverse movement is stopped and the tool is *extracted* from the material, leaving behind an exithole, i.e. a hole corresponding to the tool probe. During the welding period the process may become stationary in the sense that the temperature field and material flow, as seen from the tool, do not change. Typical values for the welding speed and rotational speed are $u_w \leq 10\,mm/s$ and $\omega \leq 1500\,RPM$. The local velocity of a point on the tool shoulder edge is determined by the rotational speed, the shoulder radius $R_{shoulder}$ and the translational welding velocity. For typical tool dimensions with $R_{shoulder}$ in the order of $10\,mm$ the rotational effect on the local velocity is much greater than the translational velocity effect.

The work in this thesis is focused on the *welding* part of the process. During this phase the workpiece material is heated due to the rotation of the tool and then stirred by the probe such that material from the two plates merges and creates the weld.

FSW is well suited for welding aluminium and the majority of work presented in the literature focuses on welding of different aluminium alloys. Yet, also materials like steel, Zhu and Chao (2004), and copper, Russell et al. (2001), or even dissimilar materials, Shercliff et al. (2005), may be welded with FSW. Industrial use of FSW is found in the marine, aerospace, railroad, and automotive industries where joining of aluminium parts are used, see TWI (2007) for a list of fields of application.

FSW has a number of advantages compared to conventional welding methods. One is the fact that there is no melting of the material, such that there is no need for inert gasses to prevent the molten material from reacting with the air. This reduces the com-

plexity of the welding setup and reduces environmental impact. On the other hand a major disadvantage is the need for efficient clamping equipment and a suitable backingplate due to the large forces involved (in Colegrove et al. (2003) downward forces up to $50\,kN$ and traversing forces up to $12\,kN$ are shown). This limits the mobility of the process and makes it challenging to do welding on very large parts.

The research presented in this thesis aims at applying optimization methods to FSW process models. In the literature limited work has been presented on this topic and most work has focused on modelling different aspect of the process, ranging from analytical models describing the temperature field to 3D numerical thermomechanical models and residual stress models. The present work focuses on optimization of thermal models of varying complexity and the next section is devoted to different thermal models.

1.1 Thermal models

Thermal aspects play an important role in the modelling and understanding of FSW. While the real process is thermomechanical in the sense that the thermal and mechanical aspects of the process are coupled, a purely thermal model may still give important knowledge about FSW and can be used as the first part of, for example, an uncoupled residual stress model or a microstructure model.

Thermal models differ greatly in complexity, from analytical Rosenthal models, Rosenthal (1946), to 3D numerical models, Chen and Kovacevic (2003), but all are based on the heat conduction equation

$$\rho c \frac{\partial T}{dt} = \nabla(k \nabla T) + q \tag{1.1}$$

or in the case of an Eulerian formulation with a convective term

$$\rho c \frac{\partial T}{dt} = \nabla(k \nabla T) + q - \rho c \mathbf{u} \nabla T \tag{1.2}$$

with suitable boundary and initial conditions. In eqs. 1.1 and 1.2 ρ denotes the material density $[\frac{kg}{m^3}]$, c the heat capacity $[\frac{J}{kgK}]$, T the temperature $[K]$, k the (possibly anisotropic) conductivity $[\frac{W}{mK}]$, q the volume heat input $[\frac{W}{m^3}]$ and $\mathbf{u}$ the welding velocity vector $[\frac{m}{s}]$. The solution procedure and its complexity depends very much on the assumptions made regarding welding setup, geometry, boundary conditions and the type of heat source. Also a distinction between constant or temperature dependent material parameters is important, the latter, e.g., removing the possibility of using analytical Rosenthal solutions.

1.1.1 Heat source models

An important factor when solving the heat conduction equation for a FSW process is the q term, or an equivalent surface flux, that determines the heat input. It is a common approach to model the heat input as a surface flux rather than a volume source and a number of different heat source models are presented below. A distinction can be made between heat generated by Coulomb friction and heat generated by the plastic deformation of material during welding. In Schmidt et al. (2004); Schmidt and Hattel (2004, 2005c) the contact condition between the tool and the workpiece is used to develop expressions for the heat generation. Given the contact pressure p, $[\frac{N}{m^2}]$, between

the tool and the workpiece, the shear stress τ, $[\frac{N}{m^2}]$, between the two parts is given by

$$\tau = \mu p \tag{1.3}$$

where μ is the friction coefficient. It should be noted that both μ and p may be non-uniform across the tool-workpiece interface. If the yield shear stress of the workpiece material is larger than the contact shear stress the two surfaces will slide against each other. This is called the *sliding* condition. If, on the other hand, the contact shear stress is larger than the yield shear stress of the material it will stick to the tool and rotate with it, and the *sticking* condition exists. Also a combination of the two are possible, if the yield shear stress is smaller than the contact shear stress for small shear rates but not for larger shear rates. In that situation the material will rotate with the tool but at a slower rate. The contact variable δ is defined as

$$\delta = \frac{v_{matrix}}{v_{tool}} \tag{1.4}$$

where v_{matrix} and v_{tool} is the local velocity of the workpiece material and the tool, respectively, and $0 \leq \delta \leq 1$. For pure sliding where the material is not rotating but only elastically deformed $\delta = 0$ and for pure sticking where the material moves with the same speed as the tool $\delta = 1$. In Schmidt et al. (2004) it is argued, based on experimental data, that at least partial sticking is present.

Based on the contact condition the local heat generation per area can be written as, Schmidt and Hattel (2005c, 2008)

$$q(r) = \omega r \left(\delta \tau_{yield} + (1 - \delta)\tau_{friction}\right) \tag{1.5}$$

where, for $\delta = 1$, the total heat generation is due to plastic deformations in the workpiece material and for $\delta = 0$ the heat generation is due only to frictional heat with the friction shear stress $\tau_{friction} = \mu p$. ω is the rotational speed and $r = r(x, y)$ is the radial distance from the axis of rotation. For direct use of eq. (1.5) one should have knowledge about the (non-uniform) distribution of δ as well as the shear yield stress τ_{yield} and $\tau_{friction}$. A number of models have been presented in the literature where, instead, the total heat input Q_{total} is prescribed and subsequently used to develop expressions for the local heat input as function of position. Q_{total} may for example be estimated from experiments.

A decoupled 3D thermo-mechanical model is presented in Chao and Qi (1998) in which the thermal problem is solved before calculating the residual stresses and the distortion of the welded plate. The heat input is generated by the sliding between the tool and the shoulder. In order to account for the heat generation from the probe, the probe diameter $d_{probe} = 0$, thus effectively moving the heat generation from the probe tip to the shoulder. The radially dependent heat input per area $q(r)$, $[\frac{W}{m^2}]$, is given by the expression

$$q(r) = \frac{3Q_{total}r}{2\pi(r_0^3 - r_i^3)} \tag{1.6}$$

where Q_{total} is the total heat input, $[W]$, r_0 is the outer radius of the shoulder and r_i the inner radius of the shoulder, i.e. in this case $r_i = 0$. Eq. (1.6) is derived and shown in a slightly different form in, among others, Schmidt et al. (2004). For simplicity only heat generation from a flat shoulder with no probe is considered here. The heat generation from a small segment of the shoulder at the distance r from the axis of rotation, see Figure 1.2, is

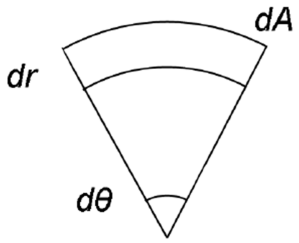

Figure 1.2: Surface element of size $dA = rd\theta dr$ used for calculation of total heat input.

$$dQ(r) = \omega r \tau dA = \omega \tau r^2 d\theta dr \tag{1.7}$$

where ω is the rotational speed and τ is the contact shear stress between the tool and the workpiece material. The surface segment dA gives a torque contribution of $r\tau dA$. Next, the total heat input can be found by integration over the shoulder area

$$\begin{aligned} Q_{total} &= \int_0^{2\pi} \int_{r_i}^{r_o} \omega \tau r^2 dr d\theta \\ &= \frac{2}{3}\pi\tau\omega(r_o^3 - r_i^3) \end{aligned} \tag{1.8}$$

In the case of sliding, as in Chao and Qi (1998), the contact shear stress is given by Coulombs law, i.e. $\tau = \tau_{friction} = \mu p$. Assuming that the total downward force P on the tool results in a uniform pressure distribution p, and that the coefficient of friction μ is also uniform, the contact shear stress is

$$\tau = \mu p = \mu \frac{P}{A_{shoulder}} = \mu \frac{P}{\pi(r_o^2 - r_i^2)} \tag{1.9}$$

For a given point the heat input per area is

$$\frac{Q}{A}(r) = \omega r \tau \tag{1.10}$$

Isolating τ in eq. (1.8) and inserting in eq. (1.10) gives eq. (1.6). Thus the heat generation per area depends linearly on the radial distance to the tool center, and the total heat input Q_{total} that may be assumed or estimated from experiments, e.g., through torque measurements. An example is given in Dickerson et al. (2003), where it is used that $Q_{total} = \omega T_{measured}$ and $T_{measured}$ is the measured torque. Alternatively, the pressure p and the friction coefficient μ may be assumed or somehow estimated such that substitution of $\tau = \mu p$ into eq. (1.8) gives

$$q = \frac{2}{3}\pi\omega r_o^3 \mu p \tag{1.11}$$

for $r_i = 0$. This may be rewritten as in Frigaard et al. (2001) to

$$q = \frac{4}{3}\pi^2 p N r_o^3 \tag{1.12}$$

where $N = \frac{\omega}{2\pi}$ is the number of revolutions per second.

In Khandkar and Khan (2001) a 3D transient model of an overlap FSW process is investigated. The heat generation is due to the sliding between the tool shoulder/probe and the workpiece material as well as the plastic deformation of the material near the probe. The total heat generation from the probe is set to 3% of the heat generation from the shoulder. The heat generation is a function of the downward force on the tool, the friction coefficient, rotational speed and is linearly dependent on the distance from the tool center. Also included in the model is a convective term to take the material transport around the probe into account. In Khandkar et al. (2003) the heat generation per area, q, is modelled as

$$q(r) = \omega r \tau \tag{1.13}$$

where τ is determined from experimental measurements of the torque during welding. A uniform value of $\tau = 14\,MPa$ is used for the alloy Al6061-T651 and therefore no assumptions on the value of δ are made. This heat generation was used in a 3D model and applied at the three contact zones between the tool and the workpiece, i.e. at the shoulder, the side of the probe and the tip of the probe.

As stated above, the FSW process is coupled in the sense that the heat generation is determined by the mechanical behaviour that in turn depends on the temperature field, thus requiring a thermomechanical model. The idea behind the so-called thermal-pseudo-mechanical (TPM) model, Schmidt and Hattel (2008), is that the friction shear stress is equal to the yield shear stress. This is the case if sticking is present, i.e. $\delta > 0$. The heat generation in eq. (1.5) is in that case

$$q = \omega r \tau_{yield} \tag{1.14}$$

This result is obtained by integration of the *volume* heat generation due to the plastic deformation of the material in a shear layer between the tool and stationary matrix material. The material in contact with the tool has the velocity $\delta\omega r$ which together with an assumption of constant strain rate in the shear layer means that the integral of the local plastic heat generation is $q_{pl} = \delta\omega r\tau_y$. This is the first part of eq. (1.5). Further, if sticking is present, i.e. $\delta > 0$, then $\tau_{friction} = \tau_{yield}$ and eq. (1.5) reduces to eq. (1.14). Generally, τ_{yield} is a decreasing function of the temperature going to zero when $T = T_{solidus}$ where the material changes from solid to liquid phase. This means that when the temperature reaches the solidus temperature for the material the heat generation vanishes, thereby limiting the obtainable maximum temperature. The model thus includes some of the mechanical effects of FSW without solving explicitly for them.

An analysis of different ways of modelling the heat source and its geometry is given in Schmidt and Hattel (2004, 2005c) where six cases are considered. The cases are: a) All the heat is generated by the shoulder with no contribution from the probe, b) Heat generation from the shoulder and a volumetric contribution from the probe and c) Heat generation from the shoulder and from the probe surface, with the probe material left out. All three models are solved assuming sliding and sticking, respectively, giving a total of six different cases. In case the sticking condition is assumed the heat is applied as a volume source in a narrow shear layer of thickness $0.5\,mm$. One of the conclusions is that the temperature field under the shoulder is greatly affected by the modelling choice whereas the farfield temperature fields are almost identical. Thus a detailed heat source model may be needed for studying effects close to the tool while a simpler model may be adequate for studying effects far from the tool.

Example with different heat sources
In this example three different heat source models are tested using a 2D Eulerian model, see Section 1.1.3. The heat source models used are the TPM model, eq. (1.14), a radially dependent model, eq. (1.6) with $r_i = 0$, and a model with uniform heat input across the tool surface. No rotational effects are included in the example, meaning that the welding velocity vector is given by $\{u_w \ 0 \ 0\}^T$. The three cases are implemented and solved in Comsol Multiphysics, Comsol (2006). Figure 1.3 shows the corresponding temperature fields. The total heat input was obtained from the TPM model by integration of eq. (1.14) and this is then applied as the prescribed total heat input in the radially dependent heat source, eq. (1.6), and in the constant heat source model. The TPM model yields a temperature field that is close to constant under the tool and a maximum temperature well below the other two models, and just below the solidus temperature set to $T_{solidus} = 805K$ in this example. The other two models predict much higher maximum temperatures and much less uniform temperatures under the tool as is clear from the temperature plots at the source. The near constant temperature in the TPM model is due to the temperature dependent yield stress used. This effectively turns off the heat source at temperatures close to the solidus temperature, and in that way limits the maximum temperature. The other two models predict almost similar temperature fields and it is noted that the maximum temperature for the radially dependent heat source is located further away from the center of the source compared to the constant heat source, where the maximum temperature is obtained only slightly behind the tool center. The local heat input close to the center is small for the radially dependent heat source and therefore the maximum temperature is obtained further back. Also note that the temperature fields are symmetric as no rotational effects are included, meaning that the material flows through the heat source parallel to the welding direction.

1.1.2 Analytical models

The Rosenthal solutions, Rosenthal (1946), are analytical equations giving the temperature field caused by an area, a line or a point heat source of strength q in one, two and three dimensions, respectively. From a FSW modelling point of view only the 2D and 3D solutions are of interest as they can be used to describe the temperature field around the tool. The solutions are obtained under a number of assumptions. The process is assumed to be steady state meaning that $\frac{\partial T}{\partial t} = 0$, the material properties k, ρ, and c are independent of the temperature and the heat source moves at a constant velocity u_w. For the 2D solution the heat flow is assumed to be in an infinitely large plane containing the welding direction and the heat source is given by a line perpendicular to the plane, i.e. q has the unit of $\frac{W}{m}$. This leads to the equation

$$\begin{aligned} T &= T_0 + \frac{q}{2\pi k} \exp(-\lambda u_w \xi) K_0(\lambda u_w r) \\ \lambda &= \frac{c}{2k} \\ r &= \sqrt{\xi^2 + y^2} \end{aligned} \tag{1.15}$$

where T_0 is the initial temperature, K_0 is the modified Bessel function of the second kind and zero'th order, ξ and y are the coordinates of the point of interest relative to the heat source, see Figure 1.4. It is noted that for $r \to 0$ the temperature $T \to \infty$ and therefore the model does obviously not predict a maximum temperature below the melting temperature as is the case in a real FSW process. The 2D Rosenthal solution

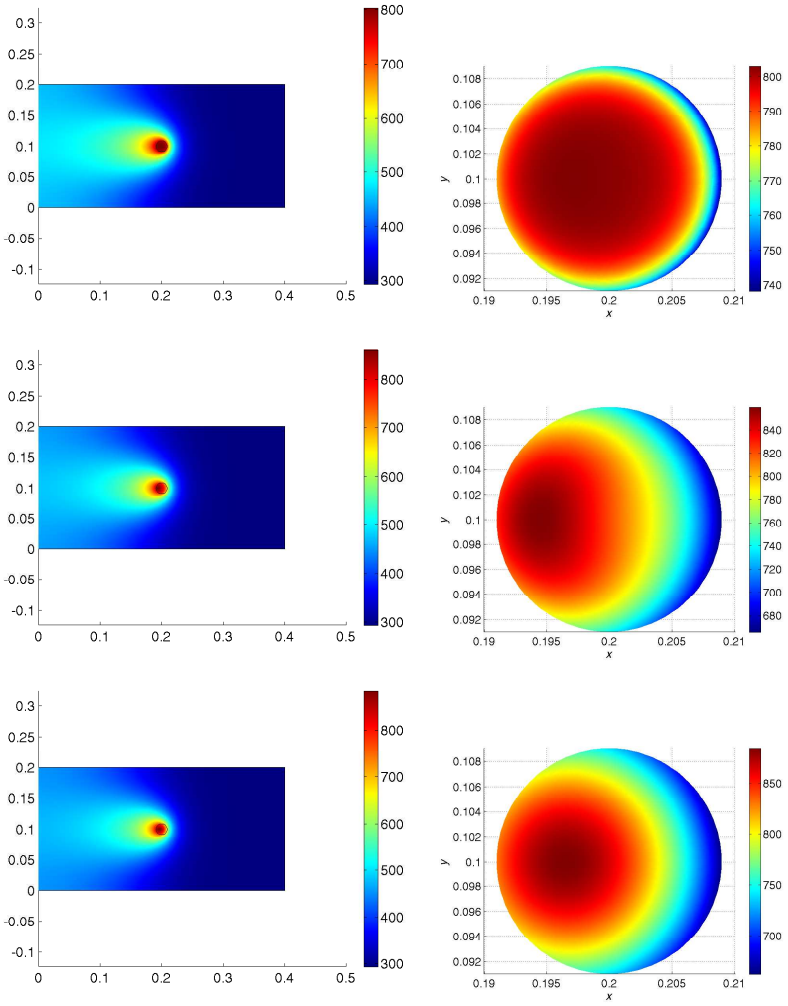

Figure 1.3: Three different heat source models. The colorbars indicate the temperature $[K]$. From the top: The TPM model (eq. (1.14), $T_{max} = 803\,K$), the radially dependent heat source (eq. (1.6), $T_{max} = 860\,K$) and the constant heat source ($T_{max} = 883\,K$). The figures to the left show global temperature fields while the right plots show the temperatures at the source. The total heat input is the same in all three examples. Notice the difference in maximum temperature. $T_{solidus}$ in the TPM model was set to $805\,K$.

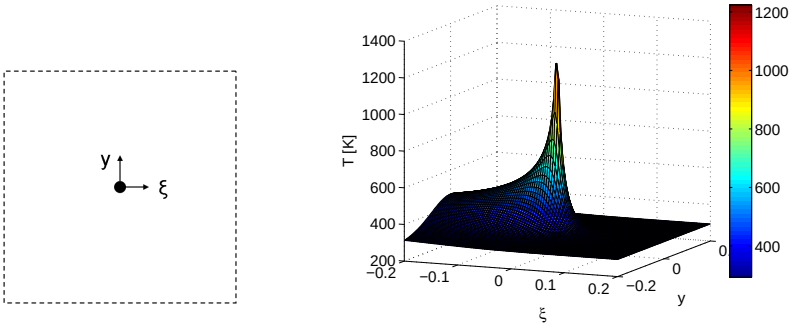

Figure 1.4: Left: Coordinates in the 2D Rosenthal solution, eq. (1.15). The welding direction is in the positive ξ direction. Dotted lines are used to indicate that the domain is infinitely large. Right: Temperature field obtained from the 2D Rosenthal solution. Note that the temperature approaches infinity at the heat source.

may be modified to take surface heat loss due to convection into account.

Similarly the solution for the 3D case, in which the heat source is modelled as a point, q has the unit W, in an infinitely large domain is given by

$$\begin{aligned} T &= T_0 + \frac{q}{4\pi k}\exp(-\lambda u_w \xi)\frac{\exp(-\lambda u_w R)}{R} \\ \lambda &= \frac{c}{2k} \\ R &= \sqrt{\xi^2 + y^2 + z^2} \end{aligned} \tag{1.16}$$

Again, as $R \to 0$ the temperature $T \to \infty$.

The two solutions shown here may be extended to take finite width (the 2D case) or finite thickness of the domain into account through the use of the method of images, Rosenthal (1946). This method results in the expressions for the temperatures being given by infinite series, in which sources at ever increasing distances are added to the solution, see Rosenthal (1946) for details.

The Rosenthal solutions described above have been widely used, especially in the early modelling of FSW. In Gould and Feng (1998) and McClure et al. (1998) the 3D Rosenthal solution is used to develop a circular heat source resembling the shoulder of the tool by placing sources in a ring around the tool center and integrating to obtain the full temperature field. The heat is assumed to be generated by Coulomb friction between the tool and the workpiece, i.e. $\delta = 0$.

In Fonda and Lambrakos (2002) the 3D Rosenthal solution taking finite thickness into account was used in an inverse modelling technique to model the temperature field in FSW. By distributing a number of point sources in the heat affected zone around the tool and scaling the contribution from each one according to the local relative speed an arbitrary temperature field was obtained. Based on experimental welds and hardness measurements the maximum temperature during welding, in a point, was estimated and used to scale the strength of the heat sources in the model.

Although the ability of the Rosenthal solutions to model the complex heat source and workpiece geometry is limited, they have the advantage of being computationally very fast compared to numerical methods. In the present thesis, this is exploited in

Articles 1 and 2 where the 2D Rosenthal solution is used as a coarse model to assist the optimization of a more expensive finite element model using space and manifold mapping techniques, see Sections 2.1.2 and 2.1.3.

1.1.3 Numerical models

In order to model the thermal aspects of FSW more accurately, numerical methods must be used, for instance finite element or finite difference methods. Numerical methods allow a more detailed geometric representation of the welding setup in which advanced tool shapes as well as the interaction between tool, workpiece and backingplate can be modelled. Also temperature dependent material parameters, like the conductivity and yield stress, can be handled by numerical methods in contrast to the analytical Rosenthal equations.

A large number of modelling choices must be made when developing a FSW model, see e.g. the overview paper by Schmidt and Hattel (2008). The possibilities include dimension (2D/3D), transient vs. steady state and Eulerian vs. Lagrangian models, i.e. whether the material flows through the mesh or whether the material follows the mesh. In this section different aspects of the modelling are presented along with some models from the literature.

FSW is a non-stationary process as described earlier with the plunge, dwell, welding, and extraction phases, and to model the full process a time-dependent model should be used, i.e. by solving eq. (1.1). However, in many cases it is reasonable to assume that the welding phase, when the tool is moving at constant velocity and not too close to the start and end points, is stationary, such that the temperature field around the tool does not change during this phase. In Schmidt et al. (2004) experimental measurements of torque and plunge force during a welding experiment are shown, and it is found that the values become almost constant during the welding phase which backs the assumption of stationarity. The present thesis focuses on thermal models of the stationary phase using an Eulerian framework. The heat equation to be solved is

$$\begin{aligned} \nabla(-k\nabla T) + \rho c_p \mathbf{u}\nabla T &= q \quad \text{in} \quad \Omega \\ T &= T_0 \quad \text{on} \quad \Gamma_T \\ -k\nabla T \cdot \mathbf{n} &= q_R \quad \text{on} \quad \Gamma_q \end{aligned} \tag{1.17}$$

where Ω is the computational domain, Γ_T is a Dirichlet boundary with given temperature T_0 and Γ_q is a Neumann boundary where the heat flux q_R is given. $\mathbf{n}$ is an outward pointing unit vector normal to the boundary. For an insulating boundary $q_R = 0$ and for boundaries in contact with e.g. air or the backingplate $q_R = h(T - T_A)$ with T_A being the temperature of the neighbouring body or fluid. For a surface heat input q_R may be given by one of the expressions in Section 1.1.1.

This type of model has a number of advantages compared to transient Lagrangian models when only the stationary response is of interest. First of all it is fast, as just one solution is required and secondly, the heat source is stationary relative to the mesh meaning that a very fine mesh can be used in the vicinity of the heat source while using a coarser mesh further away. This type of model is however normally not used for modelling more advanced effects, such as residual stresses, where transient Lagrangian models are dominant in the literature, for example in Richards et al. (2008); Tutum et al. (2009).

An effect that has a large impact on the global temperature fields is the heat loss

from the workpiece to the backingplate governed by the equation

$$q_R = h(T - T_{bp}) \tag{1.18}$$

where h is the heat transfer coefficient and T_{bp} the temperature of the backingplate. The heat transfer coefficient is non-uniform across the area of contact and is a function of contact pressure, backingplate material, surface roughness and other factors, thus making it hard to determine appropriate values. Most of the work presented in the literature assumes a uniform h across the contact area. Examples of values used are: $h = 700\frac{W}{m^2K}$ in Schmidt and Hattel (2008), $h = 350\frac{W}{m^2K}$ in Chao et al. (2003) and $h = 5000\frac{W}{m^2K}$ in Khandkar et al. (2006). In Soundararajan et al. (2005) an attempt is made to relate the heat transfer coefficient to the contact pressure between workpiece and backingplate by calculating the contact pressure for an assumed h value and then assigning a pressure-dependent h-value based on the calculated pressure distribution. They define four zones around the tool in which h varies from $h = 3000 - 4000\frac{W}{m^2K}$ below the tool to $h = 30 - 300\frac{W}{m^2K}$ in areas close to the workpiece edges in front of the tool. In Khandkar et al. (2003) different models with and without backingplates are discussed and different heat transfer coeffiecients are applied in the model and compared to experimental temperature measurements. Uniform values of $h = 10000\frac{W}{m^2K}$, $h = 5000\frac{W}{m^2K}$ and $h = 1000\frac{W}{m^2K}$ are used. Also a non-uniform h that take the value $h = 100000\frac{W}{m^2K}$ under the shoulder, $h = 10000\frac{W}{m^2K}$ in areas previously covered by the shoulder and $h = 1000\frac{W}{m^2K}$ in other areas is tested. It is found that the non-uniform heat transfer coefficient was too large under the tool and predicted too low temperatures.

Example of different heat transfer coefficients
This example is intended to show the effect of the workpiece-backingplate heat transfer coefficient h. A 3D stationary model consisting of the workpiece and a steel backingplate was implemented in Comsol Multiphysics, Comsol (2006), with the heat source given by the TPM model, eq. (1.14). Using the TPM model means that the heat input is not the same in the three cases presented and, therefore, the effect of changing h is less readily seen than when using a fixed total heat input. Figure 1.5 shows the workpiece temperature field for three different h values.

The heat transfer coefficient has a clear influence on the temperature fields by removing heat from the workpiece. This is most obvious on the left column of plots where it can be seen how the high temperature regions become more and more localised around the tool as h is increased. The effect of h is studied in more detail in Article 4 where optimization techniques and experimental data are used in order to find optimal values of h that minimize the difference between the calculated and the measured temperatures.

Although most models in the literature deal primarily with the workpiece, and in some cases the backingplate, some models also take the tool into account. As well as heat losses to the backingplate, heat is lost to the tool and a welding efficiency can be defined as the ratio of heat that is conducted into the tool to the total mechanical power, Dickerson et al. (2003). In that paper values of around 10% were determined for two different welding situations. Similarly in Khandkar et al. (2003) a value of 13% is mentioned.

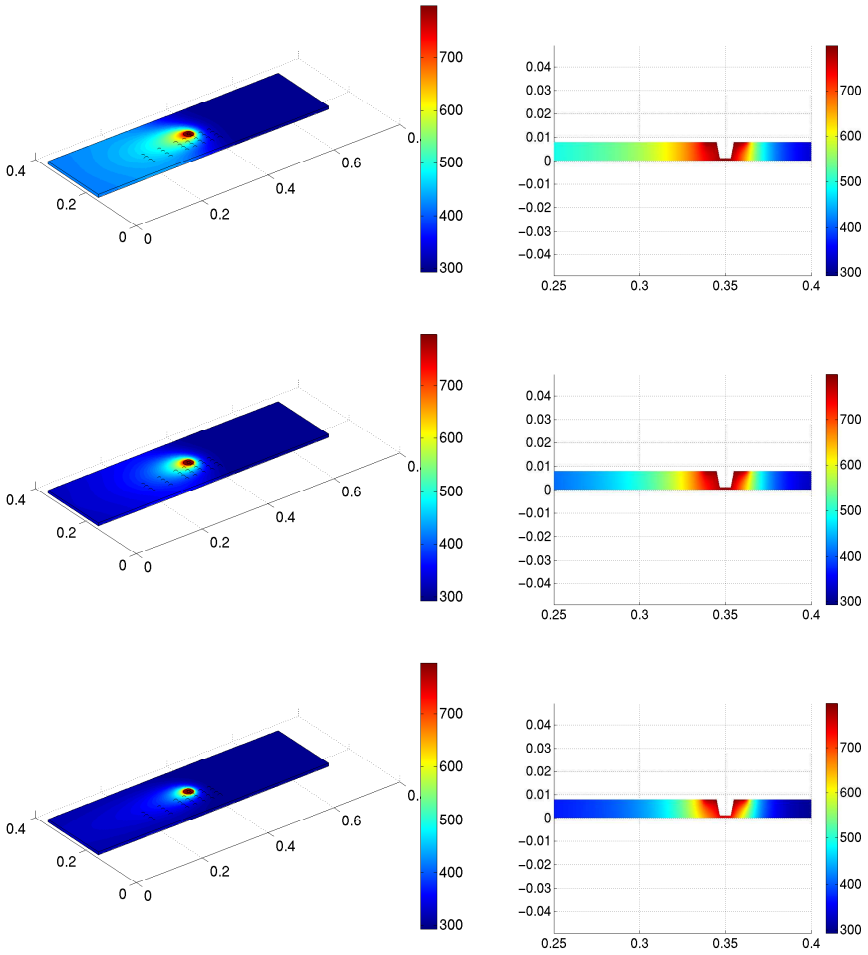

Figure 1.5: The figure shows the effect of the workpiece-backingplate heat transfer coefficient h for a 3D model of the workpiece and the backingplate. The left column shows the temperature field in the workpiece while the right column shows a cross section of the workpiece at the tool. The top row shows results for $h = 50\ \frac{W}{m^2K}$, the middle row shows results for $h = 400\ \frac{W}{m^2K}$ and the last row shows results for $h = 10000\ \frac{W}{m^2K}$. Notice the probehole and how the temperature decreases when moving away from the tool for increasing h values.

1.2 Mechanical and material aspects of FSW

The thermal analysis plays an important role in modelling of FSW and is a necessary part of more advanced models of the mechanical and material response of the workpiece. This section gives examples of some of the different models described in the literature.

1.2.1 Residual stresses and flow models

Residual stresses are important for the use of the welded material and several papers deal with the subject. Most typically a transient Lagrangian model is used for calculation of the thermal and mechanical responses. In case only residual stresses from the volumetric expansion of the material due to the heating are considered, and not the complex deformation of material around the tool, the analysis may be decoupled. This means that a purely thermal problem is solved first and afterwards the calculated temperature field is used as input to a mechanical model.

An early attempt to calculate residual stresses in friction stir welded plates is presented in Chao and Qi (1998) where a decoupled thermal and thermomechanical finite element method is used. A temperature dependent yield strength of the material is used based on the peak temperature experienced during the heating of the material. The maximum residual stresses are about 30% of the yield strength of the original material. Chen and Kovacevic (2003) study the same material, Al 6061-T6, and find residual stresses of similar magnitude. Both papers show results before and after release of the clamping and find that the residual stresses change significantly after the release, in terms of magnitude and direction. Khandkar et al. (2006) find that the magnitude of the longitudinal residual stresses are around the yield strength of the material at room temperature for two different aluminium alloys, AA-2024 and AA-6061. In Richards et al. (2008) a residual stress calculation is performed using a decoupled thermo-mechanical analysis where the tool is modelled simply as a heat source given by eq. (1.6) and thus no mechanical effect of the tool is taken into account. The yield stress of the material play an important role when calculating residual stresses, and the effect of microstructural changes is modelled using a softening model adopted from Myhr and Grong (1991a), see also Section 1.2.2. Besides calculation of residual stresses the paper deals with tensioning techniques in which the plate is loaded in tension during the welding process. This is reported to efficiently reduce the longitudinal residual stresses.

The material flow around the probe during FSW may be studied using computational solid mechanics models or computational fluid dynamics models. An example of the former is presented in Schmidt and Hattel (2005b) where the material flow, including void formation, is modelled using a local 3D model of the workpiece. The model is coupled such that the temperature field as well as the material deformation is part of the solution. Computational fluid dynamics models may also be used to study the material flow around the probe during friction stir welding. The use of this type of models are motivated by the fact that the temperatures close to the tool during welding are high, such that the material is softened substantially. An example of a CFD model is presented in Colegrove and Shercliff (2003) where a 2D model is used to calculate the planar flow of material around the tool. Two different boundary conditions are applied at the tool-workpiece interface; a sticking and a slip condition and it is found that the velocity fields around the tool and the torque and traversing force are quite different for the two models. Four non-circular tool shapes are tested but it is found that, although very different, the shape does almost not affect the torque and traversing force. A sec-

ond example of a 2D CFD model is presented in Schmidt and Hattel (2005a) where a circular tool is examined and used for describing different regions in the flow field. In this model a sticking boundary condition is used on the tool-workpiece interface. Neither of the two papers solve for the temperature fields.

1.2.2 Material hardness

While maximum temperatures during FSW are low compared to those seen in conventional welding methods, some alloys still experience a decrease in hardness HV and strength in the heat affected zone (HAZ) close to the tool due to the thermal impact. This effect is the focus of Article 3 in which the process parameters are optimized in order to obtain a given hardness distribution in the welded plate.

In Myhr and Grong (1991a) a model for determining the hardness of a 6082-T6 aluminium alloy after an arbitrary thermal treatment, such as welding, is presented. In the T6 heat treated condition this alloy has high values of hardness and strength due to fine $\mathrm{Mg_2Si}$ precipitates and the model uses the relative fraction $1 - X_d$ of these particles to interpolate between the maximum and the minimum possible hardness of the material. X_d is known to decrease as a function of time and temperature, such that a heat treatment at an elevated temperature will decrease X_d and thereby the hardness. Expressions for calculation of X_d for thermal cycles such as welding are presented. This is based on calculation of the (temperature dependent) time t_1^* required to reach full dissolution of the strengthening particles such that integration over the full thermal cycle gives the total effect on X_d. In equation form X_d is determined by

$$X_d^{1/n_1} = \int_0^{t_s} \frac{1}{t_1^*} dt \tag{1.19}$$

where t is the time and t_s is the duration of the weld cycle. n_1 is a time exponent that is usually assumed constant and equal to 0.5. Instead of calculating X_d from eq. (1.19) X_d may be determined by interpolation in so-called master curves showing X_d as a function of $\int_0^{t_s} \frac{1}{t_1^*} dt$, Robson et al. (2006). t_1^* is the time required to obtain full dissolution at a given temperature T and can be determined experimentally by measuring the time required, t_{r1}^*, for the hardness to drop from HV_{max} to HV_{min}, or X_d to go from 0 to 1, at a given temperature T_{r1}. HV_{max} and HV_{min} is the base-material hardness and the fully softened material hardness, respectively. For temperatures T different from T_{r1} the equivalent time t_1^* may be calculated by

$$t_1^* = t_{r1}^* \exp\left[\left(\frac{Q_s}{n_1 R} + \frac{Q_d}{R}\right)\left(\frac{1}{T} - \frac{1}{T_{r1}}\right)\right] \tag{1.20}$$

where t_{r1}^* and T_{r1} are the reference time and temperature, respectively. For continuous heat treatments, such as welding, where the material is heated and subsequently cooled, eq. (1.19) can be used to determine X_d. Note that t_1^* depends on T such that when inserting t_1^* into eq. (1.19) integration in time of the thermal history is required to obtain X_d. The remaining symbols in eq. (1.20) are all material constants. Once X_d is determined, the Vickers hardness HV may be calculated by

$$1 - X_d = \frac{HV - HV_{min}}{HV_{max} - HV_{min}} \tag{1.21}$$

i.e. linear interpolation between the base-material hardness HV_{max} and the fully softened material hardness HV_{min}.

Part two of the article, Myhr and Grong (1991b), shows an example of the model applied to a fusion welding process using a 3D Rosenthal solution for calculation of the temperature field. Also, linear relationships between the hardness and the yield and ultimate strenght, respectively, are shown.

The model has been applied to friction stir welding in a number of articles, e.g. in Shercliff et al. (2005) where it is applied to different aluminium alloys. Also, they show calculated and experimental hardness profiles and find reasonable agreement. Finally they discuss how the hardness model may be used in an attempt to optimize the welding speed by requiring full softening, i.e. $X_d = 1$, of the material in front of the tool such that the tool wear and risk of probe failure is reduced. A trial-and-error method is used to determine the maximum allowable welding speed. In Frigaard et al. (2001) the hardness model is applied to a friction stir welding process modelled using a finite difference approach for obtaining the thermal history. Again, the calculated hardness is compared to experimental values and the results are comparable. The hardness model presented in the present work is only based on the thermal effects of the welding and not the mechanical effects taking place close to the tool. However, in Robson et al. (2006) it is claimed that the effect of the deformation is small compared to the thermal effects. Also the model does not predict the natural aging taking place in the softened zone after welding where some of the hardness is recovered.

Example of hardness after welding
In this example the hardness model presented in eqs. (1.19),(1.20) and (1.21) is applied to a 6082-T6 aluminium alloy as in the original article by Myhr and Grong (1991a). A 3D thermal finite element model is used to calculate the thermal field used as input in the hardness model. The thermal model uses the TPM model, eq. (1.14), and consists of the workpiece and the backingplate. The thermal and microstructure models were implemented and solved in Comsol Multiphysics, Comsol (2006). Figure 1.6 shows the resulting hardness profiles in a region around the tool for two different welding speeds u_w. Increasing u_w results in a narrower heat affected zone and, therefore, a narrower region of softened material as would be expected.

The hardness model presented above is the focus of Article 3 that deals with optimization of the hardness profile after welding. Also, the material softening in front of the tool is considered taking the pointwise temperature and the microstructural changes into account. Similarly this is used in Appendix E as a constraint when maximizing the advance per revolution.

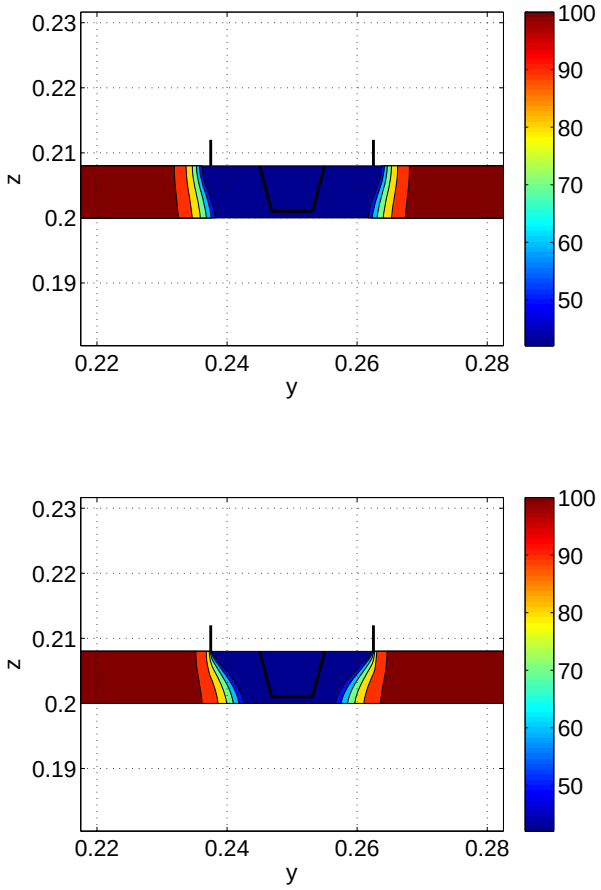

Figure 1.6: The figure shows the, calculated, hardness profile after friction stir welding of a 6082-T6 plate. Top: Welding speed $u_w = 0.5\,mm/s$. Bottom: Welding speed $u_w = 2\,mm/s$. The black lines indicate the probe volume and the shoulder radius. The color scale gives the hardness ranging from $HV_{min} = 42$ to $HV_{max} = 110$. Notice that the softened zone extends just outside the shoulder radius for the top surface while the width of the soft zone decreases towards the bottom surface. This effect is most noticable in the lower plot where the welding speed is larger and therefore the integrated thermal impact during the welding is smaller, leading to less reduction in hardness.

Chapter 2

Optimization techniques

The purpose of the present chapter is to introduce a number of optimization techniques that can be applied to optimization of FSW models. In the following sections the space- and manifold mapping techniques, that are used in Articles 1-3, are presented with the aim of developing a framework for optimization of FSW.

The optimization problems considered in this work, and in FSW in general, are characterized by having few optimization variables, in most cases the translational welding speed and the rotational speed or heat input, and relatively few constraints. The use of complex numerical models means that the computations may become very expensive and, in some cases, calculation of reliable analytic[1] sensitivities of objective and constraint functions is not possible or may be prohibited by time constraints. Thus in order to develop a general strategy applicable to a wide range of models, the optimization techniques presented here focus on the use of surrogate models that may be applied to models where gradient information is not available. Especially we focus on space and manifold mapping techniques, Bandler et al. (2004); Echeverria and Hemker (2008), where a mapping is created and iteratively updated such that when applied to a coarse model a good approximation of the true model to be optimized is obtained. Since we are dealing with mathematical models of reduced accuracy of a complex physical process we do not require very accurate solutions to the optimization problems. Instead the focus is on obtaining good, close to optimal, solutions with only a few expensive function evaluations. This means that we are not aiming at satisfying the necessary optimality conditions, the KKT conditions, see e.g. Bazaraa et al. (1993), with great accuracy. Rather, more loose convergence criteria based on relative changes in function values and optimization step lengths are used.

Mathematical optimization techniques have so far not been widely applied to friction stir welding and relatively few articles have been published on the topic. An example is given by Nandan et al. (2007) that use a genetic optimization algorithm to determine four process parameters by minimizing the differences between a numerical model and experiments. The parameters are the friction coefficient, the extent of sticking, the heat transfer coefficient at the bottom surface of the workpieces, and the amount of viscous dissipation converted into heat. The objective function to be minimized contains the peak temperature obtained in a given distance from the tool and the time the temperature is above 523 K in the given distance. Good agreement between the measured and the calculated temperatures are obtained using the optimized parameters.

[1] Meaning setting up equations to be solved for the sensitivities.

Tutum et al. (2009) used a genetic algorithm in order to solve multi-objective optimization problems considering residual stresses and tool wear. In Fratini and Corona (2007) a steepest descent method is used to maximize the joint strength of a friction stir welded lap joint using the welding speed u_w and the rotational speed ω as optimization variables. The procedure is purely experimental so gradients are obtained using a forward finite difference approach. In order to reduce uncertainties related to the experiments several identical samples were used at each set of variables and the final results show an increase in the strength after the optimization. Jayaraman et al. (2009) analyse the effect of u_w, ω, and the axial tool force on the tensile strength of the welded material. A full-factorial matrix is used to determine the points used for experiments. In Gebhard and Zaeh (2007) the authors establish an empirical relation between four parameters of the welding, including rotational speed and translational speed, by performing a number of experiments and creating a second order polynomial model for the tool temperature as function of the four parameters. The objective of the work is not optimization but modelling and understanding of the process by being able to quickly determine the effect of changes in welding parameters. In other cases trial-and-error approaches have been used in order to improve process parameters or fit calculated data to experimental data. An example of the first case is given in Shercliff et al. (2005) where the welding speed was determined such that the material in front of the tool was sufficiently softened to allow easy tool traversing. Examples of the second case include fitting the value of the workpiece-backingplate heat transfer coefficient, Khandkar et al. (2003); Schmidt and Hattel (2008), to reduce differences in calculated and measured temperatures.

The limited use of optimization techniques in FSW may be a result of the process being fairly new and that focus so far has been on obtaining good models that accurately resemble the process. Using models there is a risk that the optimizer takes advantage of the fact that some physical phenomenon is not modelled or modelled with reduced accuracy. One such example could be the problem of maximizing the welding speed where a limiting factor could be the risk of void creation in the weldline. This phenomenom can not be modelled with for example a purely thermal model and the optimizer could easily go beyond the optimization variable range where the model makes physical sense. Still, optimization techniques are a way to improve the process in a more efficient way than the trial and error methods, as mentioned above, that have been dominating so far.

This chapter deals with different aspects of the optimization methods used in this work. Two main themes are discussed; optimization using surrogate models and robust optimization. The first part is mainly concerned with space- and manifold mapping techniques whereas the second part focuses on formulations and methods for solving robust optimization problems.

A general optimization problem with equality and inequality constraints can be written as

$$\begin{aligned} \min_{\mathbf{x}} \quad & \Phi(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq 0 \quad i = 1, ..., m \\ & h_j(\mathbf{x}) = 0 \quad j = 1, ..., n \end{aligned} \tag{2.1}$$

where Φ is the objective function to be minimized, $\mathbf{x}$ are the optimization variables to be determined, g are the inequality constraints and h are the equality constraints. The problem can be solved in a number of different ways, including gradient based methods, see e.g. Gill et al. (1981), stochastic search methods like genetic algorithms,

Bäck et al. (1997), derivative free methods such as the Nelder and Mead simplex method, Nelder and Mead (1965), and surrogate model methods like response surface techniques, McDonald et al. (2007).

2.1 Surrogate techniques

The use of surrogate models in optimization is motivated by the fact that many models are so computationally expensive that traditional methods are not practical, e.g. because gradients are not easily calculated or because of time constraints. When using surrogate models, the aim is to take advantage of fast, approximate models to speed up the optimization process. Surrogate models are also known as meta models.

The accurate and expensive model that should be optimized is denoted the *high-fidelity* or *fine* model whereas the approximate model is called the *low-fidelity* or *coarse* model. In the following we denote the fine model response and the coarse model response $\mathbf{f}_f$ and $\mathbf{f}_c$, respectively, where we use bold letters to show that the responses may be vector-valued. In Robinson et al. (2006) low-fidelity models are divided into three categories

- Data fits, e.g. response surfaces.
- Reduced order models, e.g. using modal analysis.
- Hierarchical models, e.g. obtained by omitting physics modelled by the high-fidelity model.

In some situations the low-fidelity model may have a different number of optimization variables than the high-fidelity model, but this case is not considered in the present work where, in all cases, we have the same number of optimization variables with the same physical meaning in both models. Furthermore, in Robinson et al. (2006) a distinction is made between the meaning of *surrogate* model and *low-fidelity* model where the former is defined over the same design variable space as the high-fidelity model while, as mentioned, the low-fidelity model can be defined over a different design space.

When performing surrogate-based optimization, a correction may be applied to the coarse model thereby creating the surrogate model to be optimized. In the following, different types of corrections are presented. These are additive/multiplicative corrections, space mapping, in which a mapping is applied to the optimization variables, and manifold mapping where a mapping of the coarse model response is done.

2.1.1 Additive and multiplicative correction

The idea behind additive and multiplicative corrections is to do simple corrections of the coarse model response such that it approximates the fine model response more accurately. For scalar responses the exact corrections are defined as, Eldred et al. (2004),

$$\begin{aligned} A(\mathbf{x}) &= f_f(\mathbf{x}) - f_c(\mathbf{x}) \\ B(\mathbf{x}) &= \frac{f_f(\mathbf{x})}{f_c(\mathbf{x})} \end{aligned} \tag{2.2}$$

for the case of additive and multiplicative corrections, respectively. $A(\mathbf{x})$ and $B(\mathbf{x})$ are

not known but may be approximated using Taylor expansions around the current iterate $\mathbf{x}_k$

$$\begin{aligned}\alpha(\mathbf{x}) &= A(\mathbf{x}_k) + \nabla A(\mathbf{x}_k)^T(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^T\nabla^2 A(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k)\\ \beta(\mathbf{x}) &= B(\mathbf{x}_k) + \nabla B(\mathbf{x}_k)^T(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^T\nabla^2 B(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k)\end{aligned} \tag{2.3}$$

where $A(\mathbf{x}_k)$, $\nabla A(\mathbf{x}_k)$, $\nabla^2 A(\mathbf{x}_k)$, $B(\mathbf{x}_k)$, $\nabla B(\mathbf{x}_k)$ and $\nabla^2 B(\mathbf{x}_k)$ can be calculated from Eq. (2.2) and therefore contain information about the function values and first and second derivatives of both the fine and the coarse model. The corrected coarse models to be used as the surrogate model are then

$$\begin{aligned}\tilde{f}_\alpha(\mathbf{x}) &= f_c(\mathbf{x}) + \alpha(\mathbf{x})\\ \tilde{f}_\beta(\mathbf{x}) &= f_c(\mathbf{x})\beta(\mathbf{x})\end{aligned} \tag{2.4}$$

for the additive and multiplicative correction, respectively. Yet, these methods require calculation of first and second derivatives of the fine model response which could be computationally expensive. Instead the derivatives may be approximated using for example Broyden and BFGS updates, Gill et al. (1981). A slightly different approach is used in Vitali et al. (2002) where the fine and coarse models are evaluated at a number of data points and the difference and ratio measures are calculated at each data point. A response surface is then fitted to the calculated errors. In the optimization procedure the fine model can then be approximated by the coarse model response corrected using the response surface of errors.

The space and manifold mapping techniques described in the next two subsections have the advantages that they neither require the user to supply gradient information of the fine model nor that a number of fine model evaluations are performed before starting the optimization procedure.

2.1.2 Space mapping techniques

Space mapping was first introduced by Bandler et al. (1994) for optimization of electromagnetic circuits. The central idea is to do a mapping of the optimization variables in such a way that the coarse model with the mapped variables gives the same response as the fine model. Establishing this mapping, the mapped coarse model can be used as a surrogate for the fine model when solving the optimization problem. The method has gained popularity in the field of electromagnetic optimization, see for example Bakr et al. (2001); Bandler et al. (2004); Encica et al. (2007), but only relatively few papers have been published on space mapping in other fields, see e.g. Leary et al. (2001); Redhe and Nilsson (2006); Jansson et al. (2003) for examples in structural optimization. The method is general however and applicable to situations where a fine and a coarse model of the same physical problem are available.

In the following, the space mapping technique is presented for an unconstrained optimization problem in order to simplify the notation and show clearly the steps involved in updating the mapping. Constraints may be added to the problem by creating additional mappings for each constraint function. As stated earlier, the idea is to do a mapping $\mathbf{p}$ of the optimization variables such that

$$\mathbf{f}_c(\mathbf{p}(\mathbf{x})) \approx \mathbf{f}_f(\mathbf{x}). \tag{2.5}$$

In this way the fine model in the optimization problem

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}}\ \Phi(\mathbf{f}_f(\mathbf{x})) \tag{2.6}$$

can be substituted by the (simpler) mapped coarse model, i.e.

$$\mathbf{x}^{sm} = \underset{\mathbf{x}}{\operatorname{argmin}}\ \Phi(\mathbf{f}_c(\mathbf{p}(\mathbf{x}))) \tag{2.7}$$

to obtain the space mapping solution $\mathbf{x}^{sm}$. It is not guaranteed that $\mathbf{x}^{sm} = \mathbf{x}^*$ as shown in a later example. Φ is an objective function depending on the response $\mathbf{f}_f$ or $\mathbf{f}_c$, e.g. a norm. Note that the responses $\mathbf{f}_f$ and $\mathbf{f}_c$ may be scalar- or vectorvalued. As the mapping $\mathbf{p}$ is not known, space mapping is an iterative process in which the mapping is updated and the mapped coarse model is optimized once in each iteration until a satisfactory result is obtained. Although the mapping $\mathbf{p}$ is unavailable, its value for a given set of optimization variables $\mathbf{x}_k$, in iteration k, can be calculated as

$$\mathbf{p}(\mathbf{x}_k) = \mathbf{z}_k = \underset{\mathbf{z}}{\operatorname{argmin}} \|\mathbf{f}_f(\mathbf{x}_k) - \mathbf{f}_c(\mathbf{z})\| \tag{2.8}$$

where $\mathbf{z}$ are the coarse model optimization variables. In this thesis $\mathbf{x}$ and $\mathbf{z}$ have the same dimension and represent the same physical quantities and although this is the most usual situation, it is not a requirement, see e.g. Bilzer et al. (2005) where two models with a different number of optimization variables are used. Solving eq. (2.8) is known as *parameter extraction* as it gives the coarse model variables (or parameters) that minimize the difference between the fine model and the coarse model responses for the given point $\mathbf{x}_k$. If $\mathbf{p}(\mathbf{x}^*) = \mathbf{z}^*$, with $\mathbf{z}^* = \operatorname{argmin}_{\mathbf{z}}\ \Phi(\mathbf{f}_c(\mathbf{z}))$, i.e. the coarse model minimizer, the mapping is said to be perfect. For $\mathbf{p}$ to be used in the mapped coarse model optimization problem, eq. (2.7), it is approximated with a linear function around the current point $\mathbf{x}_k$ such that

$$\mathbf{p}(\mathbf{x}) = \mathbf{z}_k + \mathbf{B}(\mathbf{x} - \mathbf{x}_k) \tag{2.9}$$

where $\mathbf{B}$ is the Jacobian of the mapping function. The exact $\mathbf{B}$ is not known and therefore it is approximated using Broydens update, Broyden (1965), that is a method for sequentially approximating the Jacobian of a function based only on its function values. The update is given by

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{z}_{k+1} - \mathbf{z}_k - \mathbf{B}_k\mathbf{h}_k}{\mathbf{h}_k^T\mathbf{h}_k}\mathbf{h}_k^T \tag{2.10}$$

in which $\mathbf{h}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ is the step taken in the fine model optimization variables. In the first iteration $\mathbf{B}_0$ is equal to the identity matrix and $\mathbf{z}_0 = \mathbf{x}_0$, i.e. from eq. (2.9) $\mathbf{p}(\mathbf{x}) = \mathbf{x}$, such that the first optimization of the mapped coarse model corresponds to optimizing the coarse model directly. The iterative space mapping procedure using the Broyden update and linearization of $\mathbf{p}(\mathbf{x})$ is sometimes refered to as *aggressive* space mapping.

We note that the main optimization problem, eq. (2.7), as well as the parameter extraction problem, eq. (2.8), are solved using the coarse model. A fine model evaluation is however required in the parameter extraction in order to evaluate $\mathbf{p}$, and this will typically be the most time consuming part of a space mapping optimization. In its basic form the space mapping technique consist of the four steps shown in Figure 2.1.

Do until convergence

- Calculate the new ***fine*** model optimization variables (from the *coarse* model with mapped optimization variables):

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}}\, \Phi(\mathbf{f}_c(\mathbf{p}_k(\mathbf{x})))$$

- Evaluate the ***fine*** model:

$$\mathbf{f}_f(\mathbf{x}_{k+1})$$

- Parameter extraction.
  Calculate the value of the mapping, $\mathbf{z}_{k+1}$, at point $\mathbf{x}_{k+1}$:

$$\mathbf{z}_{k+1} = \underset{\mathbf{z}}{\operatorname{argmin}} \|\mathbf{f}_f(\mathbf{x}_{k+1}) - \mathbf{f}_c(\mathbf{z})\|$$

- Update $\mathbf{B}$

end do

Figure 2.1: The basic space mapping algorithm.

The algorithm presented above is called the dual space mapping algorithm in Echeverria and Hemker (2005) whereas the primal space mapping algorithm is obtained by changing the update of $\mathbf{x}_{k+1}$ to $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{p}_{k+1}(\mathbf{x}) - \mathbf{z}^*\|$. Similarly, in the original space mapping algorithm $\mathbf{x}_{k+1}$ is obtained by $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{h}_k$ where $\mathbf{h}_k$ is the solution to $\mathbf{B}_k\mathbf{h}_k = -(\mathbf{p}(\mathbf{x}_k) - \mathbf{z}^*)$. Only the dual space mapping algorithm has been used in this work.

Several modified versions of the basic algorithm shown above have been presented in the literature. The focus of these are often on using a trust region strategy to include a step lenght limitation in the algorithm or on different approaches to the parameter extraction. In Bakr et al. (1998) a trust region was introduced into the original scheme and in Leary et al. (2001) a trust region was included in a constrained version of the dual scheme. In that article, the trust region size is controlled by measuring the accuracy of the space mapping. If the actual decrease in the fine model is close to the predicted value, the trust region is increased and conversely if the prediction is not good enough the trust region is decreased. The constraints are handled by creating additional mappings for each of the constraint functions. In practice this type of trust region is included in the space mapping by replacing the first step in the above algorithm by

$$\begin{aligned} \mathbf{x}_{k+1} = & \underset{\mathbf{x}}{\operatorname{argmin}}\, \Phi(\mathbf{f}_c(\mathbf{p}_k(\mathbf{x}))) \\ & \text{s.t. } \|\mathbf{x} - \mathbf{x}_k\| \leq \Delta_{TR} \end{aligned} \tag{2.11}$$

with Δ_{TR} being the size of the trust region. In Redhe and Nilsson (2006) the authors use a region of interest that limits the allowable step lengths in a similar way as the trust region mentioned above is used. This region of interest is updated in a different way and only allowed to shrink during iterations, based on $\mathbf{x}_{k+1}$ compared to $\mathbf{x}_k$, i.e. the step lengths. Appendix F shows different trust region updates in more detail.

Another important aspect of space mapping is the parameter extraction step. In the basic algorithm, as presented above, the value of the mapping is established by matching the function values of the two models at a single point. This could lead to non-unique results and cause an unsuccesful space mapping optimization. A number of different methods for overcoming this risk are presented in Bandler et al. (2004). These include the multipoint method in which the fine and coarse model responses are matched at the current point $\mathbf{x}_{k+1}$ (as in the algorithm above) as well as at a number of points in the vicinity of $\mathbf{x}_{k+1}$. This method has the clear disadvantage, though,

that it requires a fine model evaluation for each additional point used and as such may become very expensive. Another method is based on matching not only function values but also the gradients of the fine and coarse models for the given $\mathbf{x}_{k+1}$. This approach is used in Leary et al. (2001) where the derivatives of the fine model are estimated by finite differences and therefore, similar to the multipoint method, requires additional fine model evaluations. It is noted that the methods requiring multiple (simultaneous) fine model evaluations does not significantly increase the computational cost if the additional fine model evaluations are performed in parallel.

Example of a space mapping
An example illustrating the effect of the space mapping is presented in Søndergaard (2003) and summarized here. The fine model considered is

$$f_f(x) = \frac{1}{2}x^2 - x + 2 \quad , \quad x \in [-2, 5]$$

and the two coarse models are

$$f_{c1}(z) = (z-1)^2 - 2(z-1) + 4 \quad , \quad z \in [-2, 5]$$
$$f_{c2}(z) = (z-1)^2 - 2(z-1) - 1 \quad , \quad z \in [-2, 5].$$

Note that $f_f(x) < f_{c1}(x) \, \forall \, x$ and that the responses are scalarvalued. The minimizers of the three functions are $x^* = 1$ and $z^* = 2$ for the fine and the two coarse models, respectively. The simple functions in this example allow the mapping to be easily calculated and the mapped coarse model is shown in Figure 2.2 along with the original coarse model and the fine model. Consider first the case of f_{c1}. The value of the

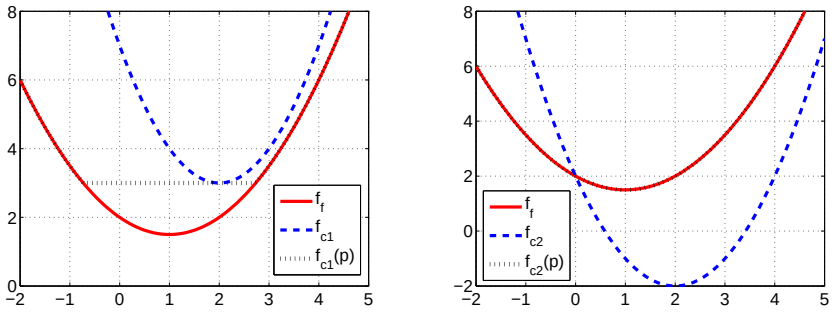

Figure 2.2: An example showing a fine model and two coarse model approximations and the space mapped coarse models. Left: Coarse model 1. Right: Coarse model 2

mapping is calculated with eq. (2.8) and we see that for all x where $f_f(x) < f_{c1}(z^*)$ the mapped coarse model gives identical responses, i.e. $f_{c1}(p(x)) = f_{c1}(z^*)$. The mapping is established using eq. (2.8) which has the solution z^* for all points x where $f_f(x) \leq f_c(z) \; \forall z$ i.e. all points x where $f_f(x)$ can not be reached by $f_{c1}(z)$ for any z. Thus minimizing $f_{c1}(p(x))$ gives no unique solution but a whole range where the true fine model optimizer $x^* = 1$ is just one of them. In this case the mapping is perfect as $p(x^*) = \mathrm{argmin}_z \, \|\mathrm{f_f(x^*) - f_{c1}(z)}\| = \mathrm{z^*}$.

The situation is different for the second coarse model where every fine model value can be reached by the coarse model. Thus $f_f(x) = f_{c2}(p(x)) \; \forall \, x$ and

$f_f(x^*) = f_{c2}(p(x^*))$. In this case however the mapping is not perfect as $p(x^*) = \mathrm{argmin}_z \|\mathrm{f_f(x^*)} - \mathrm{f_{c1}(z)}\| \neq \mathrm{z^*}$ and the original space mapping formulation of solving $p_k(x_{k+1}) = z^*$ has no solution because no x exists such that $p(x) = z^* = \mathrm{argmin_z} \|\mathrm{f_f(x)} - \mathrm{f_{c2}(z)}\|$ i.e. $f_{c2}(z^*)$ is not the best fit to $f_f(x)$ for any x.

The algorithm presented in Figure 2.1 can be modified to solve constrained optimization problems. Similarly to the unconstrained case an additional mapping $\mathbf{p}^{con}$ may be performed for each of the constraint functions g_f, such that $g_f(\mathbf{x}) \approx g_c(\mathbf{p}^{con}(\mathbf{x}))$ where the superscript con is used to distinguish the constraint mapping from the mapping used for the objective function. g_f and g_c are the fine and coarse model constraint functions, respectively. To evaluate the mapping for a given $\mathbf{x}$ another parameter extraction is performed, thereby adding an additional (unconstrained) coarse model optimization to the algorithm. In Article 1 a FSW related constrained optimization problem is solved using space mapping.

2.1.3 Manifold mapping techniques

The manifold mapping technique was introduced in Echeverria and Hemker (2005) as an extension of the space mapping technique. The extension consists of an additional mapping applied to the *response* of the coarse model such that the fine model response is approximated by

$$\mathbf{f}_f(\mathbf{x}) \approx \mathbf{S}(\mathbf{f}_c(\mathbf{p}(\mathbf{x}))). \tag{2.12}$$

For the usual case, where the number and physical meaning of the optimization variables in each of the two models are the same, the mapping $\mathbf{p}$ can be chosen as the identity mapping such that only $\mathbf{S}$ is applied to the response. The manifold mapping technique has not yet been applied to a great number of problems, as with space mapping most examples are found in the field of electromagnetics, see e.g. Echeverria et al. (2006); Lahaye et al. (2007). In Jagt (2007) it is applied to an airfoil optimization problem.

Again, the case of an unconstrained optimization problem is presented first. The fine model optimization problem is formulated as

$$\mathbf{x}^* = \underset{\mathbf{x}}{\mathrm{argmin}}\, \Phi(\mathbf{f}_f(\mathbf{x}), \mathbf{y}) = \underset{\mathbf{x}}{\mathrm{argmin}}\, \|\mathbf{f}_f(\mathbf{x}) - \mathbf{y}\| \tag{2.13}$$

where $\mathbf{f}_f(\mathbf{x})$ is the vector of fine model responses and $\mathbf{y}$ is a vector of desired values of the responses. This formulation is different from the traditional form, as in eq. (2.1), but an optimization problem may typically be reformulated to fit the form (2.13). Comparing to the space mapping formulation, eq. (2.6), the manifold mapping formulation (2.13) corresponds to choosing $\Phi(\cdot) = \|\cdot - \mathbf{y}\|$. The mapping $\mathbf{S}$ is chosen as

$$\mathbf{S}(\mathbf{v}) = \mathbf{f}_f(\mathbf{x}^*) + \bar{\mathbf{S}}(\mathbf{v} - \mathbf{f}_c(\mathbf{x}^*)) \tag{2.14}$$

where $\mathbf{v} = \mathbf{f}_c(\mathbf{x})$ is the coarse model response vector. $\bar{\mathbf{S}}$ is chosen as

$$\bar{\mathbf{S}} = \mathbf{J}_f(\mathbf{x}^*)\mathbf{J}_c^\dagger(\mathbf{x}^*) \tag{2.15}$$

where $\mathbf{J}_f$ and $\mathbf{J}_c$ are the Jacobians of the fine and coarse model, respectively, and † denote the pseudoinverse. Note that eq. (2.15) may be rewritten as $\bar{\mathbf{S}}\mathbf{J}_c(\mathbf{x}^*) = \mathbf{J}_f(\mathbf{x}^*)$ and that, in the case $\mathbf{x} = \mathbf{x}^*$, eq. (2.14) reduces to $\mathbf{S}(\mathbf{f}_c(\mathbf{x}^*)) = \mathbf{f}_f(\mathbf{x}^*)$ such that in the

optimum, the coarse model response and Jacobian is mapped to the corresponding fine model quantities. As before, the fine model is assumed to be of high complexity and to be time consuming such that $\mathbf{J}_f$ is not available, and determining $\mathbf{x}^*$ is the aim of the optimization and it is not known in advance. Therefore, the matrix $\bar{\mathbf{S}}$ is approximated during the optimization iterations. This may be done in different ways, e.g. using a Broyden update as is typically done in space mapping, or using the expressions

$$\begin{aligned}
\bar{\mathbf{S}}_{k+1} &= \Delta\mathbf{F}_f \Delta\mathbf{F}_c^{\dagger} \\
\Delta\mathbf{F}_f &= [\mathbf{f_f}(\mathbf{x}_{k+1}) - \mathbf{f_f}(\mathbf{x}_k), ..., \mathbf{f_f}(\mathbf{x}_{k+1}) - \mathbf{f_f}(\mathbf{x}_{max(k+1-n,0)})] \\
\Delta\mathbf{F}_c &= [\mathbf{f_c}(\mathbf{x}_{k+1}) - \mathbf{f_c}(\mathbf{x}_k), ..., \mathbf{f_c}(\mathbf{x}_{k+1}) - \mathbf{f_c}(\mathbf{x}_{max(k+1-n,0)})]
\end{aligned} \tag{2.16}$$

as done in most references, e.g. Echeverria and Hemker (2008).

Based on the equations presented above, the manifold mapping optimization scheme in its basic form is shown in Figure 2.3.

Do until convergence

- Calculate the new ***fine*** model optimization variables (from the *mapped coarse* model):

$$\begin{aligned}
\mathbf{x}_{k+1} &= \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{S}_k(\mathbf{f}_c(\mathbf{x})) - \mathbf{y}\| \\
&= \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{f}_f(\mathbf{x}_k) + \Delta\mathbf{F}_f \Delta\mathbf{F}_c^{\dagger}(\mathbf{f}_c(\mathbf{x}) - \mathbf{f}_c(\mathbf{x}_k)) - \mathbf{y}\| \\
&= \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{f}_c(\mathbf{x}) - \Big(\mathbf{f}_c(\mathbf{x}_k) - \Delta\mathbf{F}_c \Delta\mathbf{F}_f^{\dagger}(\mathbf{f}_f(\mathbf{x}) - \mathbf{y}\Big) \|
\end{aligned}$$

- Evaluate the ***fine*** model:

$$\mathbf{f}_f(\mathbf{x}_{k+1})$$

- Update $\Delta\mathbf{F}_f$ and $\Delta\mathbf{F}_c$.

end do

Figure 2.3: The basic manifold mapping algorithm.

Compared to the space mapping algorithm shown in Section 2.1.2 it is important to note that there is no parameter extraction step, and therefore only one optimization problem to be solved in each iteration. This property seems attractive compared to the space mapping technique where the parameter extraction requires separate optimization problems to be solved as discussed in Section 2.1.2.

In Echeverria (2007) a regularized version of manifold mapping is introduced with the aim of limiting the steps taken during the optimization iterations. From the algorithm shown above we see that if $\Delta\mathbf{F}_f$ becomes close to zero then $\Delta\mathbf{F}_f^{\dagger}$ becomes large which could lead to the term in parenthesis becoming very large such that $\mathbf{x}_{k+1}$ is very far from $\mathbf{x}_k$. This is overcome by use of a generalized singular value decomposition of the matrices $\Delta\mathbf{F}_f$ and $\Delta\mathbf{F}_c$ such that

$$\begin{aligned}
\Delta\mathbf{F}_f &= \mathbf{U}_f \mathbf{S}_f \mathbf{V}^T \\
\Delta\mathbf{F}_c &= \mathbf{U}_c \mathbf{S}_c \mathbf{V}^T.
\end{aligned} \tag{2.17}$$

Then it is possible to write

$$\begin{aligned}\Delta\mathbf{F}_c\Delta\mathbf{F}_f^{\dagger} &= \mathbf{U}_c\mathbf{S}_c\mathbf{V}^T\mathbf{V}^{-T}\mathbf{S}_f^{\dagger}\mathbf{U}_f^T \\ &= \mathbf{U}_c\mathbf{S}_c\mathbf{S}_f^{\dagger}\mathbf{U}_f^T \\ &= \mathbf{U}_c\mathrm{diag}\left(\frac{\mathbf{S}_{\mathrm{c,ii}}}{\mathbf{S}_{\mathrm{f,ii}}}\right)\mathbf{U}_{\mathrm{f}}^{\mathrm{T}}.\end{aligned} \tag{2.18}$$

In Echeverria (2007) it is suggested to replace the term $\mathrm{diag}\left(\frac{\mathbf{S}_{\mathrm{c,ii}}}{\mathbf{S}_{\mathrm{f,ii}}}\right)$ by

$$\mathrm{diag}\left(\frac{\mathbf{S}_{\mathrm{c,ii}}+\lambda\mathbf{S}_{\mathrm{c,11}}}{\mathbf{S}_{\mathrm{f,ii}}+\lambda\mathbf{S}_{\mathrm{f,11}}}\right) \tag{2.19}$$

and use λ as a regularization parameter. $\lambda = 0$ correponds to the original case while $\lambda > 0$ introduces regularization. The parameter λ is updated based on the success of the iteration in terms of reducing the fine model objective function. If the iteration is not successful λ is increased and the new point $\mathbf{x}_{k+1}$ replaced by $\mathbf{x}_{k+1} = \mathbf{x}_k + \tau(\mathbf{x}_{k+1} - \mathbf{x}_k)$ where $0 < \tau < 1$. Thus the regularization is not enforced as a standard trust region where the allowable step length is directly controlled and an unsuccessful step results in a new optimization with reduced trust region size.

In case of a constrained optimization problem additional mappings are created for the constraints, Echeverria (2007), based on the same concepts shown above. A constrained version of manifold mapping is used in Articles 2 and 3. In Article 3 the constrained manifold mapping scheme is combined with a classic trust region and filter method such that only points that reduce the objective function value or the constraint violation are accepted. Trust region and filter methods are described in more detail in Appendix F.

Beam example using space and manifold mapping
In order to illustrate the effect of the mapping in the space and manifold mapping techniques, an example of optimization of a cantilever beam is shown. Using an analytical model allows the response and the mapping to be calculated for a large number of parameter values. The response to be optimized is the deflection at the tip of the beam given by the expression $u_{tip} = \frac{PL^3}{3EI}$ where P is the load, L the length of the beam, E the modulus of elasticity and I is the moment of inertia, see Figure 2.4. In the fine model the cross section is assumed circular such that $I_f = \frac{\pi}{4}x^4$ where x is the radius of the beam and the variable to be optimized. In the coarse model the cross section is assumed rectangular with fixed width B and the moment of inertia is $I_c = \frac{1}{12}Bx^3$ with x being the height of the beam. We note that in this example the design variable x does not have the exact same physical meaning in both models but rather represents a characteristic dimension of the given cross section.

Minimizing the tip deflection has a trivial solution in which x should be as large as possible, so instead we aim at obtaining a specific value, $\tilde{u}$, of the deflection. In the case of space mapping this is done by minimizing $\Phi(u_{tip}(x)) = (utip(x) - \tilde{u})^2$ and in the case of manifold mapping directly by setting $y = \tilde{u}$ in eq. (2.13).

Figure 2.5 shows the response and the corresponding objective function value for the fine and the coarse model. We see that the response decays with increasing x and that the optima, where the objective function value is zero, for the two models are not identical.

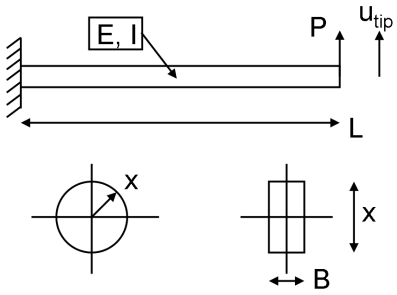

Figure 2.4: Sketch of the beam used in space and manifold mapping example. In the bottom: The two cross-sections, to the left the circular fine model cross-section and to the right the coarse model cross-section.

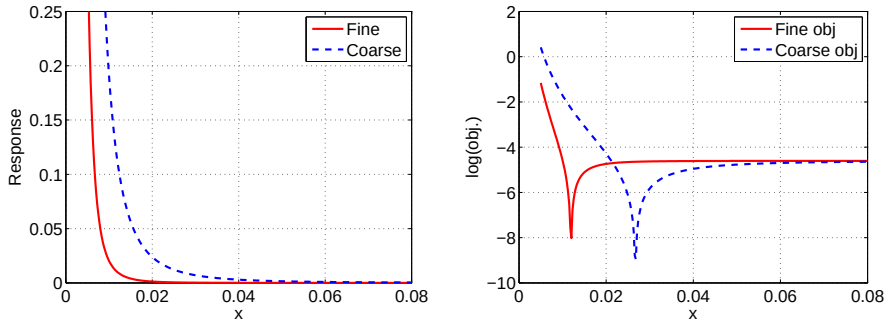

Figure 2.5: Left: The response (tip deflection) of the fine and coarse model. Right: Objective function values corresponding to the left plot.

Figure 2.6 shows results after the first iteration, i.e. the coarse model optimum, the corresponding fine model evaluation and the coarse model with the updated mapping. Note that the mapped coarse model minimizer, still, does not match the fine model minimizer after the first iteration.

Both methods converge to the point $x = 0.0119$ which is the fine model optimum. Figure 2.7 shows the fine and coarse models as well as the mapped coarse model objective functions after the last iteration. We see that the objective function based on the mapped coarse models are not identical for the two methods, but the minimum of the two are found at the same x where also the fine model has its minimum.

A common property of space and manifold mapping techniques is that they aim at *locally* mapping the coarse model such that it resembles the fine model. Thus the fine and mapped coarse model responses do not necessarily show good agreement away from the current point. In Bakr et al. (1998) the mapping available after optimization is used for Monte Carlo simulations of the response at points close to the optimum by substituting the expensive fine model evaluations by the fast mapped coarse model evaluations. This is an interesting idea that may be useful in connection to robust

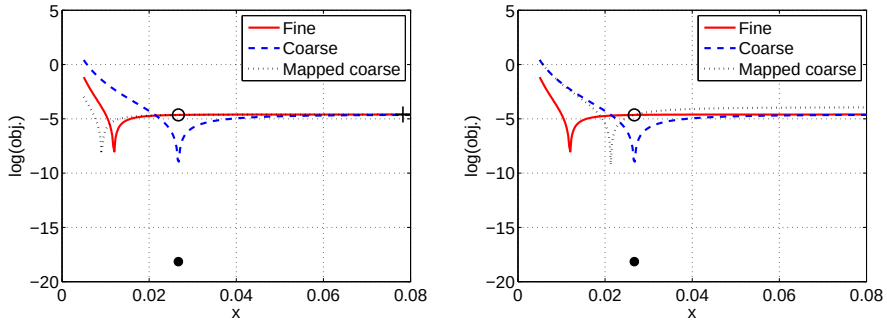

Figure 2.6: Left: Space mapping after the first iteration. The fine and coarse model objective functions are plotted along with the mapped coarse model. The black dot indicates the coarse model optimum, the black circle the fine model evaluated in the coarse model optimum and the black cross (at $x \approx 0.075$) the point obtained through parameter extraction, i.e. where the response of the two models match. Right: Similar plot for manifold mapping, but without parameter extraction. Note that a finite number of x values have been used for plotting so that the response curves do not reach $-\infty$ as they should where the objective function is zero.

optimization.

A possibly problematic situation may occur in case of constrained optimization using space and manifold mapping if there is no feasible solution to the coarse model optimization problem. In that case, the optimizer should proceed by minimizing the constraint violation such that a new $\mathbf{x}_{k+1}$ is generated and the procedure can continue.

2.1.4 Response surface techniques

The first part of this chapter dealt with surrogate optimization techniques where a correction is applied to a given coarse model. In this section focus is on a specific type of coarse models, namely response surfaces. This type of approximate model is often used as the coarse model in combination with space and manifold mapping optimization. Examples are given in Jansson et al. (2003); Redhe and Nilsson (2006); Lahaye et al. (2007) for problems in solid mechanics and electromagnetics. In the present work a response surface is applied as coarse model in a manifold mapping scheme in Article 3.

Response surface techniques are popular due to the simplicity of setting up the optimization process, and the fast evaluation once the function values at the data points are obtained. Also, gradients are easily obtained via analytical expressions and the fast evaluation times mean that it is possible to try multiple starting points for the optimization. Clearly, response surfaces may also be used for optimization without any of the above mentioned corrections as done in McDonald et al. (2007) and other articles.

In general, response surface methods are based on evaluations of the fine model in a number of data points. They therefore suffer from the curse of dimensionality, meaning that for n optimization variables a regular grid, where the range of each variable is discretized into m points, results in a total of m^n data points which may become extremely expensive for large n, i.e. many optimization variables. More advanced

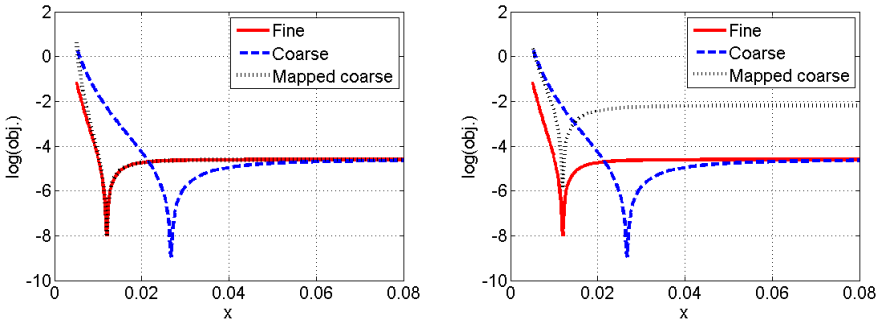

Figure 2.7: Left: Space mapping after the last iteration. The fine and coarse model objective functions are plotted along with the mapped coarse model objective function (the black dotted line). Right: Similar plot for manifold mapping.

sampling techniques based on design of experiments may help overcome (in part) this problem by selecting sampling points in an optimized way, see e.g. Crary (2002), such that the number of data points may be reduced. The problems considered in this work, however, have a small number of design variables and response surface techniques may be a useful tool for optimization.

The method used in the present work, Article 3, is a response surface based on radial basis functions, see for example McDonald et al. (2007); Krishnamurthy (2003). This type is well suited for global approximation of a function because of the ability to reproduce multiple extrema and complex, in the sense that it is not well approximated by e.g. a low-order polynomial, responses. This is in contrast to, for example, a quadratic polynomial model that has at most one interior extremum and limited ability to approximate complex functions. The approximation, $f_c(\mathbf{x})$, of the scalar fine model response $f_f(\mathbf{x})$ in a point $\mathbf{x}$ is given by

$$\begin{aligned} f_c(\mathbf{x}) &= \sum_{i=1}^{N} \lambda_i \phi(r, c) \\ r &= \|\mathbf{x} - \mathbf{x}_i\|_2 \end{aligned} \tag{2.20}$$

where N is the number of data points (number of fine model evaluations), λ_i are the coefficients, ϕ is the radial basis function, c is a constant and $\mathbf{x}_i$ are the data points. Different choices exist for the radial basis functions, see, among others, McDonald et al. (2007); Krishnamurthy (2003); Fang et al. (2005)

- Thin-plate spline: $\phi(r, c) = r^2 \log(cr^2)\,,\ 0 < c \leq 1$
- Cubic: $\phi(r, c) = (r + c)^3$
- Multiquadric: $\phi(r, c) = \sqrt{(r^2 + c^2)}$
- Inverse multiquadric: $\phi(r, c) = \frac{1}{\sqrt{(r^2+c^2)}}$
- Gaussian: $\phi(r, c) = \exp(-cr^2)\,,\ 0 < c \leq 1$

By requiring that the approximation interpolates the function values exactly at the datapoints, the coefficients λ_i can be found by solving the equation system

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{F} \tag{2.21}$$

where

$$\mathbf{A} = \begin{bmatrix} \phi(r_{11}, c) & \dots & \phi(r_{1N}, c) \\ \vdots & \ddots & \vdots \\ \phi(r_{N1}, c) & \dots & \phi(r_{NN}, c) \end{bmatrix}, \ \mathbf{F} = \begin{Bmatrix} F_1 \\ \vdots \\ F_N \end{Bmatrix}, \ \boldsymbol{\lambda} = \begin{Bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \end{Bmatrix} \tag{2.22}$$

and $r_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. The constant c is used to control the smoothness of the response surface. The value of c should be chosen such that the response surface approximates the fine model optimaly, however there is no easy guidelines for choosing c. Large values of c may cause problems as the conditioning number of $\mathbf{A}$ becomes large.

An important factor when using gradient based methods for optimization of radial basis function approximations is the existence of well defined gradients. For these to exist at datapoints, i.e. where $r_{ii} = 0$, it is required that $c \neq 0$ for the multiquadric and inverse multiquadric. Similar conditions exist for other choices of basis functions. As mentioned above, a response surface based on radial basis functions interpolates the fine function values exactly, by construction, at the data points which seems attractive for (deterministic) computer programs that give identical responses when evaluated repeatedly for fixed $\mathbf{x}$.

It is known that radial basis function approximations are not well suited for reproducing constant functions, Krishnamurthy (2003). This problem is overcome by augmenting eq. (2.20) with a number of polynomial terms such that

$$f_c(\mathbf{x}) = \sum_{i=1}^{N} \lambda_i \phi(r) + \sum_{j=1}^{M} P_j(\mathbf{x}) b_j \tag{2.23}$$

where

$$P_j = \begin{bmatrix} 1 & x & y & x^2 & xy & ... \end{bmatrix} \tag{2.24}$$

for a problem of 2 variables. In order to determine the extra M coefficients b_j it is required that $\sum_{i=1}^{N} \lambda_i P_j(\mathbf{x}_i) = 0$ for $j = 1, ..., M$. This condition results in the following equation system to be solved for the $N + M$ coefficients

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & 0 \end{bmatrix} \begin{Bmatrix} \boldsymbol{\lambda} \\ \mathbf{b} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F} \\ 0 \end{Bmatrix} \tag{2.25}$$

where $\mathbf{B}_{ij} = P_j(\mathbf{x}_i)$ and $\mathbf{b}$ contains the coefficients b_j. The coefficient matrix $\mathbf{A}$ in eq. (2.22) is dense, meaning that for a large number of datapoints it becomes computationally expensive to calculate $\boldsymbol{\lambda}$. In that case, compactly supported radial basis functions, Ho et al. (2005), may be used such that $\mathbf{A}$ becomes sparse.

Example of a radial basis function response surface

The following example shows a response surface created using the above mentioned technique. The data is the total heat input into the workpiece during FSW obtained by integration of eq. (1.14) over the tool-workpiece interface for a 3D Eulerian model similar to the one used in Article 3. The variables considered are the linear welding speed and the rotational speed and a regular grid of 5 by 5 data points is used. The

datapoints cover a very large range with $u_w \in [1, 15]\, mm/s$ and $\omega \in [30, 500]\, rad/s$ which spans the range of typical welding parameters. Figure 2.8 shows the resulting response surface for the augmented radial basis function method, eq. (2.23), using the multiquadric basis functions. Notice that the radial basis function model predicts

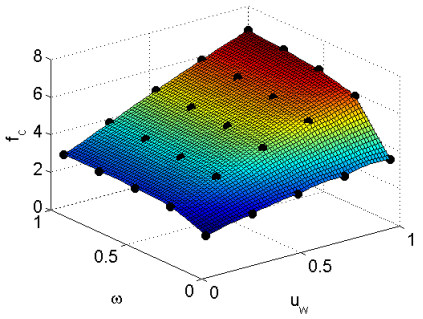

Figure 2.8: The total heat input, $[W]$, for a 3D thermal model approximated using the augmented radial basis function method, eq. (2.23). Note that the variables u_w and ω are scaled to be in the range $[0, 1]$ and that the response surface predicts the correct values at all datapoints, indicated by black dots, by construction.

the correct value at all datapoints by construction. The response seems to be almost linearly increasing in variable u_w while the ω dependence is more piecewise linear. The present example uses the augmented radial basis function method with only a constant term added but results obtained with only the radial part are similar.

Besides radial basis function methods kriging, see e.g. Lophaven et al. (2002); Quiepo et al. (2005), may be used for creating global approximation models. Similar to the radial basis function method, kriging has the advantage that it interpolates the fine model function value at data points exactly, gradients are easily obtained and it can estimate multiple extrema and complex responses.

The use of response surfaces in this work has been as coarse models in connection with space and manifold mapping techniques and focus has been on obtaining *rough* approximations of the response of the fine model. Therefore, a limited number of datapoints have been used in order to speed up the process of creating the response surface and thereby sacrificing improved accuracy, the idea being that the coarse model response will eventually be mapped to the fine response. The initial phase of evaluating the fine model at the selected datapoints may be performed very efficiently in case multiple processors are available such that the evaluations at different parameter values can be performed simultaneously.

Returning to the definitions in the beginning of Section 2.1 response surfaces belong to the type of low-fidelity models denoted *data fits*. As seen above, they have the advantages of easy implementation, fast evaluation and that gradients are easily calculated. A response surface was used as coarse model in Article 3 where also a simplified finite element model, i.e. a hierarchical type model, was used. Similarly, in Articles 1 and 2 a Rosenthal solution was used as coarse model. The disadvantage of response surfaces compared to physically based coarse models is that they may, locally, not behave as the fine model. This can be caused by complex responses, such that artificial

local extrema exist in the response surface approximation. This also means, that locally the gradients of the response surface can have opposite signs than the gradients of the fine model.

2.2 Robust optimization

In this section robust optimization techniques are discussed, motivated by the fact that solving the optimization problem (2.1) yields optimization variables that are optimal for the particular problem. However, small changes in parameters, either optimization variables or other input data, may, in some cases, lead to large changes of the response and results far from the optimized one. Typical examples are changes in modulus of elasticity in a solid mechanics problem or thermal conductivity for the problems considered in this thesis. The purpose of robust optimization strategies is to include such variations in the optimization problem in such a way that the optimal results are less sensitive to uncertainties.

Often a distinction is made between robust optimization and reliability based design optimization (RBDO). Park et al. (2006) explains the difference as: "In robust design, insensitiveness of the objective function is emphasized. In reliability design, reliability of constraints is important." An example of a RBDO problem could be the minimization of the weight of a structure subject to the constraint that the probability of exceeding the yield stress, given uncertain loads and material data, should be lower than a certain value. Neither RBDO problems or the Taguchi method, which uses design of experiments techniques for evaluating different sets of optimization parameters, are studied here.

A few examples of uncertainties were mentioned above, but other types may also be considered. Inspired by the classification in Beyer and Sendhoff (2007) uncertainties related to engineering design problems can be divided into the four groups presented below, where also the particular case of FSW is considered.

- A) Changing environmental and operating conditions.
  This category includes variations in the ambient temperature during welding, which will affect the heat losses to the surroundings, or changes in material parameters such as the thermal conductivity. These are uncertainties that are not easily reduced.

- B) Production tolerances and actuator imprecision.
  This category may include the precision of the welding machine and take e.g. variations in welding speed into account. Thus in the present case these uncertainties are related to the optimization variables.

- C) Uncertainties in the system output.
  This type of uncertainties include modelling errors arising from the use of (simplified) mathematical models of the FSW process that introduces errors compared to the real process. However, we have limited information about these errors and therefore this type of uncertainties is not considered in the present work.

- D) Feasibility uncertainties.
  This category includes constraints on the values of the optimization variables, however, for the problems at hand, it is not critical that the optimization variables are within given bounds, so this type of uncertainties is not relevant here.

Besides the distinction between the different classes, the uncertainties may be modelled as *deterministic*, *probabilistic* or *possibilistic* where deterministic means that the uncertain parameter, e.g. the thickness of a plate, can vary between given values, probabilistic means that there is a given probability that the parameter takes a certain value while possibilistic means that a possibility of some event is considered.

The uncertainties considered in this work, Article 2, are the thermal conductivity, i.e. a type A) uncertainty, and the welding speed and heat input, that are both of type B). These parameters are assumed to be probabilistic and normal distributed with known mean values and standard deviations. The probability density function for a normal distribution is

$$p(b) = \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left[-\frac{1}{2}\left(\frac{b-\mu_b}{\sigma_b}\right)^2\right] \tag{2.26}$$

where μ_b and σ_b is the mean and standard deviation, respectively, of the parameter b. The parameter b is used to denote an uncertain parameter that may include optimization variables $\mathbf{x}$ or other input data. It is noted that the normal distribution has the, in this case, unphysical property that it does not exclude negative values. Although the probability is low for the (realistic) values used in the present case, e.g., the conductivity could become negative. Using a log-normal distribution instead would overcome this risk.

The above description and classification of the uncertain parameters lead to the question of how to formulate the optimization problem and subsequently solve it, see Beyer and Sendhoff (2007); Park et al. (2006) for an overview.

Generally, the robust optimization problem is concerned with reducing both the mean value and the standard deviation of the response, and may be seen as a multiobjective optimization problem. Figure 2.9 show an example of the effect of reducing the mean value and the standard deviation of a response. Instead of treating this prob-

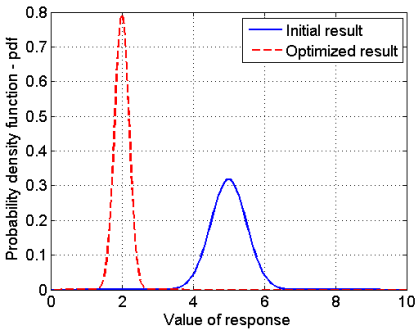

Figure 2.9: An example showing the probability density functions (pdf), eq. (2.26), for a normal distributed response for the initial data (solid blue line) and the optimum (dotted red line). In the optimum, the response has a lower mean value, corresponding to the curve being moved to the left, and lower standard deviation, meaning that the width of the curve is reduced.

lem as a multiobjective optimization problem, a common approach is to substitute the deterministic objective function Φ in problem (2.1) with a function $F(\mu_f, \sigma_f)$ of the mean value μ_f and standard deviation σ_f of the response f. Often this is formulated

as a weighted sum of μ_f and σ_f, i.e.

$$\min_{\mu_{\mathbf{x}}} \quad F = w\mu_f + (1-w)\sigma_f \tag{2.27}$$

where $0 \leq w \leq 1$. This formulation may be seen as an (unconstrained) robust version of the deterministic optimization problem in eq. (2.1). We note that this can be treated as a standard optimization problem where the uncertain effects are included in the objective function. Similarly the deterministic constraints may be substituted with robust counterparts. As the variables $\mathbf{x}$ are considered uncertain, with a given distribution, the common approach is to treat the mean value of $\mathbf{x}$, $\mu_{\mathbf{x}}$ as the optimization variable whose value is to be determined. The standard deviation $\sigma_{\mathbf{x}}$ is typically kept constant or fixed at some percentage of μ_x. However, $\sigma_{\mathbf{x}}$ may also be treated as an optimization variable while keeping in mind that a reduction in $\sigma_{\mathbf{x}}$ automatically reduces σ_f, and therefore it should be penalized in some way, to take the cost of reducing $\sigma_{\mathbf{x}}$ into account. It is noted that eq. (2.27) may be easily reformulated to suit the case where the objective is to obtain a desired mean value.

An example intended to show the possible difference between a robust and a deterministic optimum is shown in Figure 2.10. The deterministic optimum is located at

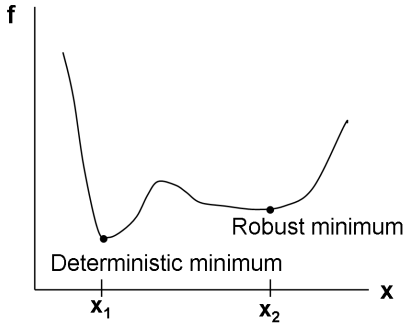

Figure 2.10: The deterministic (global) minimum is located at x_1. A small variation in x will lead to a large change in f while at x_2 a similar variation will lead to a small change in f.

$x = x_1$ where f takes its smallest value. However, if x changes slightly from this value the corresponding change in f will be large. In contrast, a small change in x at x_2 will give a small change in f, thus making $f(x_2)$ more robust to variations in x than $f(x_1)$.

Evaluating the mean and standard deviation to be used in the objective function, as in eq. (2.27), is one of the challenges in robust optimization. Using directly the integral expressions, Melchers (1987),

$$\begin{aligned} \mu_f &= \int_{-\infty}^{\infty} f(\mathbf{b}) p_b(\mathbf{b}) d\mathbf{b} \\ \sigma_f^2 &= \int_{-\infty}^{\infty} (f(\mathbf{b}) - \mu_f)^2 p_b(\mathbf{b}) d\mathbf{b} \end{aligned} \tag{2.28}$$

where p_b is the joint probability density function and $\mathbf{b}$ is a vector of n uncertain parameters, is often not possible since the integral can not be evaluated efficiently. In the present work, Article 2, only statistical independent variables have been considered,

such that $p_b = p_{b_1} \cdot p_{b_2} \cdots p_{b_n}$, Park et al. (2006). The uncertain parameters are the welding speed, the heat input and the thermal conductivity and we consider it here to be reasonable to assume that these are uncorrelated. If several material parameters, e.g. ρ, c and k, were considered as uncertain parameters it would be natural to assume that they were not independent.

Instead of using eq. (2.28) directly, it is possible to apply various approximations. A common approach is to use expressions based on Taylor expansions for the evaluation of the statistical data. In that case, the expressions for mean value and standard deviation are, Koch et al. (2004),

$$\begin{aligned} \mu_f &= f(\mu_{\mathbf{b}}) + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 f}{\partial b_i^2}(\mu_{\mathbf{b}})\,\sigma_{b_i}^2 \\ \sigma_f^2 &= \sum_{i=1}^{n}\left(\frac{\partial f}{\partial b_i}(\mu_{\mathbf{b}})\right)^2 \sigma_{b_i}^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2 f}{\partial b_i \partial b_j}(\mu_{\mathbf{b}})\right)^2 \sigma_{b_i}^2\sigma_{b_j}^2 \end{aligned} \tag{2.29}$$

for second order expansions. The use of eq. (2.29) requires second order derivative information of the response function f which in many cases is not easily obtained, meaning that often only the first order term is used, Parkinson (1995); Du and Chen (2000). In Lee et al. (2004) a combination of the first order Taylor expansions and response surface method is used, by creating local quadratic response surfaces, based on datapoints within a trust region, that in turn is used for calculation of derivatives and statistics, based on eq. (2.29). The formulation of the optimization problem is similar to eq. (2.27), and based on the success of the new point the trust region is updated and a new local quadratic response surface is build, and so forth. The fact that only function values, and not gradients, of the real function are used makes implementation easy.

Monte Carlo methods, Hammersley and Handscomb (1964), may also be used to determine the mean value and standard deviation. In Monte Carlo methods the mean and standard deviation of a function f is obtained by performing a number of calculations of the response in a number of sample points following the distributions of the uncertain parameters. Plain Monte Carlo typically require a large number of function evaluations making it too computationally expensive for engineering problems involving, for example, finite element models. Different schemes exist that allow a reduced number of evaluations, while still giving reliable results, to be used, e.g. Latin Hypercube Sampling, McKay et al. (1979). The idea behind Latin Hypercube sampling is to select the sampling points such that the sampling space is covered sufficiently accurate, even for a reduced number of points. Still, Monte Carlo methods tend to require many function evaluations thus making optimization very expensive, as the Monte Carlo simulation is inside the iterative loop of the optimization process. In Lee and Park (2006) Monte Carlo simulations are performed on a response surface, based on the real function, to evaluate the statistical data for robust optimization. This allows very fast Monte Carlo simulations that may give good results for sufficiently accurate response surfaces.

Example of robust optimization

An example illustrating the Taylor expansion and Monte Carlo approaches is shown in Figure 2.11. The objective function is similar to eq. (2.27) with the weight $w = 0.25$, i.e.

$$\min_x \quad 0.25\mu_f + (1 - 0.25)\sigma_f \tag{2.30}$$

where the response function f is a fourth order polynomial. The uncertain parameter is x, which is assumed to be normal distributed with fixed standard deviation $\sigma_x = 0.3$. Note that $w = 0.25$, meaning that much emphasis is put on reducing the standard deviation in order to show a clear difference between the robust and the deterministic optimum. Figure 2.11 shows the response function $f(x)$ as well as the objective function based on the three methods mentioned. An important point to notice is that the

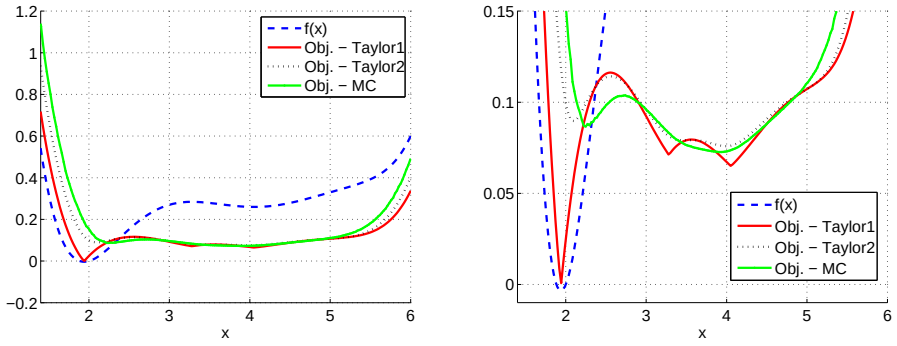

Figure 2.11: An example showing how different calculations of mean and standard deviation affect the objective function, eq. (2.30). Left: The plot shows the function $f(x)$ as well as the objective function based on first and second order Taylor expansions and Monte Carlo simulation (denoted MC). Right: Zoom of the left plot. It is clear that in this case the objective function is not well approximated by the first order Taylor expansion while the second order Taylor expansion gives results that are close to those obtained using Monte Carlo simulations.

objective function based on the first order Taylor expansion has local minima where the function $f(x)$ has stationary points, i.e. where $\frac{df}{dx} = 0$. This is expected as the standard deviation vanishes in these points, see eq. (2.29). The second order approach follows the Monte Carlo results much closer. It should be noted that in many practical cases the use of first order Taylor expansions might not be a problem. As an example, the tip deflection of a cantilever beam, as in the example in Section 2.1.3, $u_{tip} = \frac{PL^3}{3EI}$ is an increasing function of the applied load such that $\frac{du_{tip}}{dP} = \frac{L^3}{3EI} \neq 0$. Similarly, the temperature given by the Rosenthal solution, eq. (1.15), is monotone in q and u_w for a fixed position (ξ, y).

In Article 2 both first and second order expansions are used in robust optimization of a thermal model of the FSW process. It is found, for the tested uncertainties, that first order expansions give reasonable results but also, that the accuracy of the calculated mean and standard deviation is increased when using second order expansions.

Using Taylor expansions is a way of obtaining the statistical data within a reasonable time frame. This means however that use of standard gradient-based optimization methods becomes complicated by the need for higher order derivatives and the space and manifold mapping techniques constitute an efficient method for dealing with these types of problems.

Chapter 3

Summary of results

The four articles that are part of the thesis are briefly presented below. The complexity of the models and the optimization problems is increased through the articles, from a 2D thermal model in Article 1 used for studying the space and manifold mapping techniques, a 2D thermal model taking robustness into account in Article 2 and, finally, a 3D thermal model used for optimizing the hardness in Article 3. In all three articles the aim is to find optimal welding parameters, u_w and ω, that optimize the given model. Article 4 concerning the estimation of the workpiece-backingplate heat transfer coefficient h_{bottom} is a little different in the sense that the welding parameters are fixed and, instead, an inverse problem is solved in order to determine the spatial distribution of h_{bottom}.

Study of space- and manifold mapping, Article 1 (Appendix A)

Article 1 deals with an initial study of the use of space and manifold mapping techniques for solving FSW related optimization problems. The two techniques are applied to a 2D thermal optimization problem aiming at obtaining a desired temperature field in the workpiece. The methods are tested using an analytical Rosenthal model and a simplified finite element model as coarse models. Also an example is made where the fine finite element model is nonlinear due to a temperature dependent conductivity while the coarse model is the Rosenthal model with fixed conductivity. The computational cost of the methods is discussed.

Robust optimization, Article 2 (Appendix B)

Article 2 deals with robust optimization of a 2D thermal model of the FSW process. The objective is to obtain desired mean responses and reduce the standard deviation of the responses considering normal distributed uncertainties in thermal conductivity and optimization variables. The statistical data is calculated using expressions based on Taylor expansions and compared to Monte Carlo simulations. Good agreement is found.

Using derivatives of the responses in order to obtain objective and constraint function values makes it infeasible to use gradient based methods for the solution of the

optimization problems, as this would require and additional order of differentiation, and, instead, the manifold mapping technique is used with an analytical Rosenthal model as coarse model.

Optimization of hardness, Article 3 (Appendix C)

In this article the complexity of the FSW model is increased by using a 3D thermal model consisting of the workpiece and the backingplate. The heat source is given by the (nonlinear) TPM model, eq. (1.14) and the contact heat transfer between the workpiece and the backingplate is modelled by the term $h(T - T_2)$. In order to account, in the temperature field, for the flow of material around the probe, a rotational velocity field is enforced in a shear layer close to the probe. Compared to the models used in the previous two articles this model is more expensive to evaluate because of the increased number of degrees of freedom, it being a 3D model, and the nonlinearity introduced by the heat source.

The calculated temperature field is used to determine the material hardness that depends on the thermal history of a given point. The optimization problem is concerned with optimizing the hardness distribution with constraints on the total heat input and the material yield strength in front of the tool. Manifold mapping is used for solving the optimization problems using a simplified finite element model and a response surface as coarse models.

Inverse modelling of the workpiece-backingplate heat transfer coefficient, Article 4 (Appendix D)

The work presented in this article aims at determining the spatial distribution of the workpiece-backingplate heat transfer coefficient $h_{bottom}(x, y)$. Experimental temperature measurements in a number of points in the workpiece are used along with a 3D stationary finite element model of the process. The optimization problem is formulated as the minimization of the difference between the calculated and the measured temperatures. Four different parametrizations of $h_{bottom}(x, y)$ of increasing complexity are tested, and it is found that using a non-uniform $h_{bottom}(x, y)$ improves the objective function value, and thereby the model-experiment correspondance, significantly.

3.1 Contributions and impact

The work presented in this thesis focuses on applying optimization techniques to FSW models, and it should be seen as steps towards the increased use of mathematical optimization techniques for improving the FSW process through optimized welding or modelling parameters. As was shown in Chapter 2, this is an area that has not received much attention previously.

The methods used, although applied to thermal models of reduced complexity, are general and applicable to other types of models used in FSW. The manifold mapping technique with a response surface as coarse model, as is presented in Article 3, can readily be applied to optimization of other FSW models because of the generality of the method, and the fact that it does not require the process model to be modified in any way.

The results obtained using the space and manifold mapping optimization techniques may not be directly transferred to a real welding experiment but, using more advanced models, the methods provide a means for optimizing process parameters that can improve the process. This is useful especially for complex models where intuition may be limited and extensive parameter studies are not feasible.

Analyzing and optimizing the robustness of results in FSW is another field studied in this thesis. For process models many uncertainties may play a role, and the effect of variations of the parameters is not always obvious. Using the Taylor expansion based robustness measures in combination with the (derivative free) manifold mapping technique as in Article 2 seem to be an effective way of optimizing for robustness.

In Article 4 the workpiece-backingplate heat transfer is studied. To the authors knowledge this is the first use of gradient-based optimization techniques for determining the position-dependent workpiece-backingplate heat transfer coefficient. In contrast, previous work in the literature has focused on uniform heat transfer coefficients determined by genetic algorithms or trial-and-error methods. The results obtained here can be transferred to existing models without much effort in order to improve the agreement between models and experiments. As the temperature field is of great importance for the properties of the welded plate, e.g. residual stresses and microstructure, using accurate values for the heat transfer coefficient may increase the accuracy of such models.

Space mapping, and to some extend manifold mapping, techniques have gained much popularity in the field of optimization of electromagnetic components, but have been used very little in optimization of mechanical systems, examples include Redhe and Nilsson (2006); Jansson et al. (2003), where the problem types are often similar, i.e. an expensive numerical model that should be optimized. It is believed that the methods presented here may be applied succesfully to many mechanical engineering problems. An example is the use of response surfaces for optimization of complex models, where an extension of the optimization scheme with a space mapping or manifold mapping loop will not introduce many complications or a large increase in the computational cost and should improve the accuracy of the optimized results.

Chapter 4

Concluding remarks

The work presented in this thesis has focused on optimization of a FSW process with most emphasis on finding optimal welding parameters for different models. FSW involves multiple physical phenomena, making detailed modelling complex and time consuming. The models considered here are thermal models of increasing complexity. The work should be seen as a proof-of-concept in which different techniques suitable for optimization of FSW models are studied.

The use of complex analysis models makes optimization of FSW challenging due to the time required for each evaluation of the model. This makes use of surrogate model optimization techniques interesting as these methods may help to reduce the number of function evaluations required, and furthermore, do not necessarily require the user to supply gradient information related to the complex model. The last property is very appealing. The surrogate techniques used here are the space and the manifold mapping techniques, that work by establishing a mapping between the coarse model and the fine model in such a way that the fine model optimization problem is substituted by a sequence of mapped coarse model optimization problems. The methods performed well when compared to traditional optimization methods.

Three different types of coarse models have been used in this work: An analytical model, a reduced finite element model, and finally a response surface model. The first one has the advantage of being computationally very fast and of being based on the same physics, although modelled at a lower accuracy, as the fine model. The reduced finite element model is very similar to the fine model but suffer from the fact that it in itself is quite expensive to optimize. The response surface model is very fast once the fine model has been evaluated at the datapoints, which can be a time consuming process.

The space and manifold mapping techniques differ in the way the mapping is applied and the way it is updated. In space mapping the mapping is performed on the optimization variables while in manifold mapping it is performed on the coarse model response. For very inexpensive coarse models the computational cost of the two models is similar, but the need for parameter extraction, i.e. additional coarse model optimization problems, in space mapping means that for more expensive coarse models a manifold mapping iteration will be faster than a space mapping iteration.

The two techniques have been applied to thermal optimization problems in which the objective was to obtain a desired temperature field. The manifold mapping technique was furthermore applied to different robust optimization problems in which the aim was twofold: obtain a desired mean response while reducing the standard deviation

of the responses. This problem was solved using a Taylor expansion based technique for determining the statistical data for the responses based on given statistical data for the uncertain parameters. The use of manifold mapping meant that no gradients of the objective and constraint functions were required, a great advantage when using the Taylor expansion based robustness models. Still, gradients of the response functions should be used in the Taylor expansions used for calculation of the statistical data that in turn is used in the objective and constraint functions. Thus the additional order of differentiation required for gradient based optimization is avoided by use of manifold mapping.

The workpiece-backingplate heat transfer coefficient was studied through an inverse modelling approach aiming at minimizing the difference between experimental temperature measurements and calculated temperatures. A non-uniform heat transfer coefficient having a large value in region under the tool was found to perform significantly better than a uniform distribution.

4.1 Future work

The optimization techniques presented here have performed well on the problems at hand. It would however be interesting to apply them to more advanced models that more accurately model the FSW process, and in that way use the methods for determining optimal welding parameters that in turn may be used in actual experiments or production situations. The more complex models could include residual stress or flow models.

In addition to optimizing the translational and rotational speed, the tool shape could be included in the optimization problems. This requires, first of all, a model that capture the influence of the tool shape, e.g. a flow model, and secondly, application of shape or topology optimization techniques. This is likely to be a extremely computationally expensive process. However, simpler sizing problems, e.g. determining the optimal probe diameter, may also be used to improve the process.

In connection to the calculation of the workpiece-backingplate heat transfer coefficient it would be interesting to do additional experiments with more measuring points such that a larger amount of data would be available for the inverse modelling. Possibly, design of experiments techniques can be used to determine where the additional measuring points should be placed.

Bibliography

Alexandrov, N., Dennis, J., Lewis, R. and Torczon, V. 1998. A trust-region framework for managing the use of approximation models in optimization. *Structural optimization* **15**, 16–23.

Bäck, T., Hammel, U. and Schwefel, H.-P. 1997. Evolutionary computation: Comments on the history and current state. *IEEE Transactions on evolutionary computation* **1**(1), 3–17.

Bakr, M., Bandler, J., Biernacki, R., Chen, S. and Madsen, K. 1998. A trust region aggressive space mapping algorithm for em optimization. *IEEE transactions on microwave theory and techniques* **46**(12), 2412–2425.

Bakr, M., Bandler, J., Madsen, K. and Søndergaard, J. 2001. An introduction to the space mapping technique. *Optimization and engineering* **2**, 369–384.

Bandler, J., Biernacki, R., Chen, S., Grobelny, P. and Hemmers, R. 1994. Space mapping technique for electromagnetic optimization. *IEEE Transactions on microwave theory and techniques* **42, no. 12**, 2536–2544.

Bandler, J., Cheng, Q., Dakroury, S., Mohamed, A., Bakr, M., Madsen, K. and Søndergaard, J. 2004. Space mapping: The state of the art. *IEEE Transaction on microwave theory and techniques* **52, no. 1**, 337–361.

Bazaraa, M., Sherali, H. and Shetty, C.: 1993. *Nonlinear programming theory and algorithms - second edition*. John Wiley & Sons, Inc.

Beyer, H. and Sendhoff, B. 2007. Robust optimization - a comprehensive survey. *Computer methods in applied mechanics and engineering* **197**, 3190–3218.

Bilzer, H., Frank, F. and Menzel, W. 2005. A space mapping method allowing models with different parameter rank and physical meanings for coarse and fine model. *IEEE MTT-S International Microwave Symposium Digest* pp. 1127–1130.

Broyden, C. 1965. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation* **19, No. 92**, 577–593.

Chao, Y. and Qi, X. 1998. Thermal and thermo-mechanical modeling of friction stir welding of aluminum alloy 6061-t6. *Journal of Materials Processing & Manufacturing Science* **7**, 215–233.

Chao, Y., Qi, X. and Tang, W. 2003. Heat transfer in friction stir welding - experimental and numerical studies. *Journal of manufacturing science and engineering* **125**, 138–145.

Chen, C. and Kovacevic, R. 2003. Finite element modeling of friction stir welding - thermal and thermomechanical analysis. *International Journal of Machine Tools and Manufacture* **43**, 1319–1326.

Colegrove, P. and Shercliff, H. 2003. 2-dimensional CFD modeling of flow round profiled FSW tooling. *Friction stir welding and processing II, The minerals, metals & materials society* pp. 13–22.

Colegrove, P., Shercliff, H. and Threadgill, P. 2003. Modelling and development of the trivex(tm) friction stir welding tool. *4th International Symposium on Friction Stir Welding*.

Comsol: 2006. www.comsol.com.

Crary, S. 2002. Design of computer experiments for metamodel generation. *Analog integrated circuits and signal processing* **32**, 7–16.

Dickerson, T., Shi, Q. and Shercliff, H. 2003. Heat flow into friction stir welding tools. *4th International symposium on friction stir welding*.

Du, X. and Chen, W. 2000. Towards a better understanding of modelling feasibility robustness in engineering design. *Journal of Mechanical Design* **122**, 385–394.

Echeverria, D. 2007. Two new variants of the manifold-mapping technique. *COMPEL The International Journal for Computation and Mathematics in Electrical and Electronic Engineering* **26, No.2**, 334–344.

Echeverria, D. and Hemker, P. 2005. Space mapping and defect correction. *Computational methods in applied mathematics* **5, no. 2**, 107–136.

Echeverria, D. and Hemker, P. 2008. Manifold mapping: a two-level optimization technique. *Accepted for publication in Computing and Visualization in Science*.

Echeverria, D., Lahaye, D., Encica, L., Lomonova, E., Hemker, P. and Vandenput, A. 2006. Manifold-mapping optimization applied to linear actuator design. *IEEE transactions on magnetics* **42**(4), 1183–1186.

Eldred, M., Giunta, A. and Collis, S. 2004. Second-order corrections for surrogate-based optimization with model hierarchies. *AIAA paper 2004-4457*.

Encica, L., Echeverria, D., Vendenput, A., Hemker, P. and Lahaye, D. 2007. Efficient optimal design of electromagnetic actuators using space mapping. *Structural and multidisciplinary optimization* **33**, 481–491.

Fang, H., Rais-Rohani, M., Liu, Z. and Horstemeyer, M. 2005. A comparative study of metamodelling methods for multiobjective crashworthiness optimization. *Computers and structures* **83**, 2121–2136.

Fletcher, R., Leyffer, S. and Toint, P. 2002. On the global convergence of a filter-sqp algorithm. *SIAM journal on optimization* **13**(1), 44–59.

Fonda, R. and Lambrakos, S. 2002. Analysis of friction stir welds using an inverse problem approach. *Science and Technology of Welding and Joining* **7**(3), 177–181.

Fratini, L. and Corona, V. 2007. Friction stir welding lap joint resistance optimization through gradient techniques. *Journal of manufacturing science and engineering* **129**, 985–990.

Frigaard, O., Grong, O. and Midling, O. 2001. A process model for friction stir welding of age hardening aluminium alloys. *Metallurgical and materials transactions A* **32A**, 1189–1200.

Gebhard, P. and Zaeh, M. 2007. Empirical model for the tool shoulder temperature during friction stir welding. *Technische Universitt Mnchen,Institute for Machine Tools and Industrial Management*.

Gill, P., Murray, W. and Saunders, M.: 2006. Users guide for snopt version 7: Software for large-scale nonlinear programming.

Gill, P., Murray, W. and Wright, M.: 1981. *Practical optimization*. Academic Press.

Gould, J. and Feng, Z. 1998. Heat flow model for friction stir welding of aluminum alloys. *Journal of Materials Processing & Manufacturing Science* **7**, 185–194.

Hammersley, J. and Handscomb, D.: 1964. *Monte Carlo methods*. Fletcher & Son Ltd, Norwich.

Ho, S., Yang, S., Ni, G. and Wong, H. 2005. A response surface methodology based on improved compactly supported radial basis function and its application to rapid optimizations of electromagnetic devices. *IEEE Transactions on magnetics* **41**(6), 2111–2117.

Jagt, M. 2007. Airfoil optimization by using the manifold mapping technique. *Report MAS-E0709, CWI Netherlands*.

Jansson, T., Nilsson, L. and Redhe, M. 2003. Using surrogate models and response surfaces in structural optimization - with application to crashworthiness design and sheet metal forming. *Struct. Multidisc. Optim.* **25**, 129–140.

Jayaraman, M., Sivasubramanian, R., Balasubramanian, V. and Lakshminarayanan, A. K. 2009. Optimization of process parameters for friction stir welding of cast aluminium alloy A319 by taguchi method. *Journal of scientific & industrial research* **68**, 36–43.

Khandkar, M. and Khan, J. 2001. Thermal modeling of overlap friction stir welding for al-alloys. *Journal of Materials Processing and Manufacturing Science* **10**, 91–105.

Khandkar, M., Khan, J. and Reynolds, A. 2003. Input torque based thermal model of friction stir welding of al-6061. *6th International trends in welding research conference proceedings, 15-19 April 2002, Pine Mountain, GA*.

Khandkar, M., Khan, J., Reynolds, A. and Sutton, M. 2006. Predicting residual stresses in friction stir welded metals. *Journal of materials processing technology* **174**, 195–203.

Koch, P., Yang, R.-J. and Gu, L. 2004. Design for six sigma through robust optimization. *Struct. Multidisc. Optim.* **26**, 235–248.

Krishnamurthy, T. 2003. Response surface approximation with augmented and compactly supported radial basis functions.. *44th AIAA/ASME/ASCE/AHS Structures, Structural Dynamics, and Materials Conference, 7-10 April 2003, Norfolk, Virginia*.

Lahaye, D., Canova, A., Grusso, G. and Repetto, M. 2007. Adaptive manifold-mapping using multiquadric interpolation applied to linear actuator design. *COMPEL The International Journal for Computation and Mathematics in Electrical and Electronic Engineering* **26, no. 2**, 225–235.

Leary, S., Bhaskar, A. and Keane, A. 2001. A constraint mapping approach to the structural optimization of an expensive model using surrogates. *Optimization and Engineering* **2**, 385–398.

Lee, K. and Park, G. 2006. A global robust optimization using kriging based approximation model. *JSME International Journal, series C*.

Lee, Y., Hong, K.-J. and Choi, D.-H. 2004. An efficient robust optimal design method for engineering systems with numerical noise. *10th AIAA/ISSMO multidisciplinary analysis and optimization conference* **3**, 1931–1938.

Lophaven, S., Nielsen, H. and Søndergaard, J. 2002. DACE A Matlab Kriging Toolbox. *Technical Report IMM-TR-2002-12, IMM, Tech. University of Denmark*.

Matlab: 2006. The mathworks, inc.

McClure, J., Tang, W., Murr, L., Guo, X., Feng, Z. and Gould, J. 1998. A thermal model of friction stir welding. *Trends in welding research, June 1998, Pine Mountain* pp. 590–594.

McDonald, D., Grantham, W., Tabor, W. and Murphy, M. 2007. Global and local optimization using radial basis function response surface models. *Applied Mathematical Modelling* **31**, 2095–2110.

McKay, M., Beckman, R. and Canover, W. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21, No. 2**, 239–245.

Melchers, R.: 1987. *Structural Reliability Analysis and Prediction*. Ellis Horwood Ltd.

Myhr, O. and Grong, Ø. 1991a. Process modelling applied to 6082-T6 aluminium weldments-I. reaction kinetics. *Acta metall. mater.* **39**(11), 2693–2702.

Myhr, O. and Grong, Ø. 1991b. Process modelling applied to 6082-T6 aluminium weldments-II. applications of model. *Acta metall. mater.* **39**(11), 2703–2708.

Nandan, R., Prabu, B., De, A. and Debroy, T. 2007. Improving reliability of heat transfer and materials flow calculations during friction stir welding of dissimilar aluminum alloys. *Welding journal* **86**, 313–322.

Nelder, J. and Mead, R. 1965. A simplex method for function minimization. *Computer journal* **7**, 308–313.

Park, G.-J., Lee, T.-H., Lee, K. and Hwang, K.-H. 2006. Robust design: An overview. *AIAA Journal* **44, No. 1**, 181–191.

Parkinson, A. 1995. Robust mechanical design using engineering models. *Transactions of the ASME* **117**, 48–54.

Quiepo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, R. and Tucker, P. 2005. Surrogate-based analysis and optimization. *Progress in aerospace sciences* **41**, 1–28.

Redhe, M. and Nilsson, L. 2006. A multipoint version of space mapping optimization applied to vehicle crashworthiness design. *Struct. Multidisc. Optim.* **31**, 134–146.

Richards, D., Prangnell, P., Williams, S. and Withers, P. 2008. Global mechanical tensioning for the management of residual stresses in welds. *Materials science and engineering A* **489**, 351–362.

Robinson, T., Eldred, M., Willcox, K. and Haimes, R. 2006. Strategies for multifidelity optimization with variable dimensional hierarchical models. *47th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference, 1-4 May 2006, Newport, Rhode Island*.

Robson, J., Kamp, N., Sullivan, A. and Shercliff, H. 2006. Modelling precipitate evolution during friction stir welding of aerospace aluminium alloys. *Materials science forum* **519-521**, 1101–1106.

Rosenthal, D. 1946. The theory of moving sources of heat and its application to metal treatments. *Transactions of the A.S.M.E.*

Russell, M., Shercliff, H. and Threadgill, P. 2001. Development and application of an analytical process model for friction stir welding. *Aluminum 2001 - Proceedings of the TMS 2001* pp. 225–234.

Schmidt, H. and Hattel, J. 2004. Heat source models in simulation of heat flow in friction stir welding. *International Journal of Offshore and Polar Engineering* **14**(4), 296–204.

Schmidt, H. and Hattel, J. 2005a. CFD modelling of the shear layer around the tool probe in friction stir welding. *Friction Stir Welding and Processing III, San Francisco, USA, TMS*.

Schmidt, H. and Hattel, J. 2005b. A local model for the thermomechanical conditions in friction stir welding. *Modelling and simulation in materials science and engineering* **Vol. 13**, 77–93.

Schmidt, H. and Hattel, J. 2005c. Modelling heat flow around tool probe in friction stir welding. *Science and Technology of Welding and Joining.*

Schmidt, H. and Hattel, J. 2008. Thermal modelling of friction stir welding. *Scripta Materialia* **58**, 332–337.

Schmidt, H., Hattel, J. and Wert, J. 2004. An analytical model for the heat generation in friction stir welding. *Modelling and Simulation in Materials Science and Engineering* **12**, 143–157.

Shercliff, H., Russell, M., Taylor, A. and Dickerson, T. 2005. Microstructural modelling in friction stir welding of 2000 series aluminium alloys. *Mécanique & Industries* **6**, 25–35.

Søndergaard, J.: 2003. *Optimization using surrogate models - by the space mapping technique*. PhD thesis. Technical University of Denmark.

Soundararajan, V., Zekovic, S. and Kovacevic, R. 2005. Thermo-mechanical model with adaptive boundary conditions for friction stir welding of al 6061. *International journal of machine tools & manufacture* **45**, 1577–1587.

Tutum, C., Schmidt, H., Hattel, J. and Bendsøe, M. 2009. A preliminary study on optimization of residual stresses in friction stir welding. *submitted*.

TWI: 2007. The Welding Institute, www.twi.co.uk.

Vitali, R., Haftka, R. and Sankar, B. 2002. Multi-fidelity design of stiffened composit panel with a crack. *Structural and multidisciplinary optimization* **23**, 347–356.

Walmag, J. and Delhez, E. 2005. A note on trust-region radius update. *SIAM Journal on Optimization* **16, No. 2**, 548–562.

Zhu, X. and Chao, Y. 2004. Numerical simulation of transient temperature and residual stresses in friction stir welding of 304l stainless steel. *Journal of materials processing technology* **146**, 263–272.

Appendix A

Article 1

Larsen, Anders A.; Bendsøe, Martin; Hattel, Jesper; Schmidt, Henrik. Optimization of friction stir welding using space mapping and manifold mapping-an initial study of thermal aspects *Journal of Structural and Multidisciplinary Optimization* **38**(3), 289-299, DOI 10.1007/s00158-008-0288-6

Struct Multidisc Optim (2009) 38:289–299
DOI 10.1007/s00158-008-0288-6

INDUSTRIAL APPLICATION

Optimization of friction stir welding using space mapping and manifold mapping—an initial study of thermal aspects

Anders A. Larsen · Martin Bendsøe · Jesper Hattel · Henrik Schmidt

Received: 5 November 2007 / Revised: 21 March 2008 / Accepted: 19 May 2008 / Published online: 21 August 2008

Abstract The aim of this paper is to optimize a thermal model of a friction stir welding process by finding optimal welding parameters. The optimization is performed using space mapping and manifold mapping techniques in which a coarse model is used along with the fine model to be optimized. Different coarse models are applied and the results and computation time are compared to gradient based optimization using the full model. It is found that the use of space and manifold mapping reduces the computational cost significantly due to the fact that fewer function evaluations and no fine model gradient information is required.

Keywords Friction stir welding · Space mapping · Manifold mapping

1 Introduction

In this paper we find optimal welding parameters in order to obtain a specified temperature field in friction stir welding (FSW) of thin plates using a simplified thermal model. FSW is a solid state welding process in which a rotating tool is moved along the welding line. A part of the tool is submerged into the material and this causes the material to heaten and merge together to form a weld without melting the material.[1] Compared to traditional welding methods like arc welding the temperatures during FSW are low. This gives the process a number of advantages; for instance, residual stresses and distortions are smaller. Also it is well suited for welding aluminium and other materials that are hard to weld using conventional methods. Today FSW is being used in the marine, automotive and aerospace industries.

The physics involved in FSW include heat transfer, contact phenomena and large plastic deformations and this makes detailed simulations very time consuming. A number of different models are used, e.g. thermal models (Schmidt et al. 2004; Khandkar and Khan 2001), CFD models (Colegrove et al. 2003; Schmidt and Hattel 2005a) and thermomechanical models (Chen and Kovacevic 2003; Schmidt and Hattel 2005b). In this work we focus on the thermal aspects of the process as this is of importance for e.g. the microstructure and residual stresses. Part of the objective is to study the use of space and manifold mapping as a tool for optimization of FSW.

Space mapping was introduced by Bandler et al. (1994) as a method for optimization of computationally expensive problems in electromagnetics, but it is a general technique for handling design optimization of systems where one can generate models of variable complexity (see, e.g., Bakr et al. (2001) and comments

Part of this work was presented at WCSMO7, Seoul, Korea.

A. A. Larsen (✉) · M. Bendsøe
Department of Mathematics, Matematiktorvet,
Technical University of Denmark,
2800 Lyngby, Denmark
e-mail: a.a.larsen@mat.dtu.dk

J. Hattel · H. Schmidt
Department of Mechanical Engineering,
Technical University of Denmark,
2800 Lyngby, Denmark

[1] It is currently being debatted if localized melting occurs during welding.

below). Later the method has been developed further including use of trust regions and different parameter extraction schemes and it has since been applied to a large number of problems in the field of electromagnetics and microwaves. The main idea is to use two models for the optimization, one accurate and one faster but less accurate model. For details and surveys see e.g. Bandler et al. (2004) and Bakr et al. (2001). In Leary et al. (2001) space mapping is used for solving an optimization problem in solid mechanics. The problem is constrained so an extra mapping is used for the constraint. Redhe and Nilsson (2006) and Jansson et al. (2003) applied space mapping to complex problems in sheet metal forming and crashworthiness optimization. Manifold mapping, Echeverria and Hemker (2005), is a recently developed related type of method that like space mapping uses two models for optimization. Examples of application include optimization of linear actuators, see Lahaye et al. (2007).

The present paper focuses on optimization of a thermal FE model of the FSW process. The optimization problem is solved using space and manifold mapping techniques, and the results are encouraging in the sense that the number of function evaluations and the computation time is reduced compared to gradient based optimization using the full model. The paper is organized as follows: In Sections 2 and 3 the space and manifold mapping techniques are presented. Section 4 describes the optimization problem and the thermal models describing the process. Two different examples are shown in Sections 5 and 6 while Section 7 concludes the paper.

2 Space mapping

Numerical models of FSW taking many physical effects into account tend to become very complex and computationally expensive. This causes two problems from an optimization perspective; computation time may become too long for practical use and secondly it may not be possible to calculate reliable gradient information.

Space mapping is a method in which two different models are used; one computationally expensive that is to be optimized and a simpler but faster model. The two models are denoted the fine and the coarse model respectively. Both models include the same number of design variables with the same physical meaning but the coarse model is less accurate. A number of types of coarse models can be used, e.g. an analytical model, a response surface, a coarse-mesh FE model or others.

The equality constrained optimization problem to be solved has the form

$$\begin{aligned} \mathbf{x}^* = \arg\min_{\mathbf{x}} \quad & f_f(\mathbf{x}) \\ \text{s.t.} \quad & c_f(\mathbf{x}) = 0 \end{aligned} \tag{1}$$

where $\mathbf{x}^*$ are the optimal design variables, f_f is the fine model objective function, c_f is a fine model constraint function and $\mathbf{x}$ are the fine model design variables. It is assumed that gradients of the fine model are not available. The coarse model, $f_c(\mathbf{z})$ and $c_c(\mathbf{z})$, describes the same physical problem but is less expensive and gradients are assumed to be available. The design variables $\mathbf{x}$ and $\mathbf{z}$ have the same dimension.

The idea is to do a mapping $\mathbf{p}$ of the fine model design variables in such a way that the coarse model response, based on the mapped fine model design variables (i.e., $f_c(\mathbf{p}(\mathbf{x}))$, $c_c(\mathbf{p}^{\text{con}}(\mathbf{x}))$), will be similar to the fine model response. Thereby the fine, and expensive, model in the optimization problem can be replaced by a surrogate model, i.e. the coarse model with the mapped fine model variables. The principle is sketched in Fig. 1.

The mapping, however, is unknown, but should be chosen such that

$$f_c(\mathbf{p}(\mathbf{x})) \simeq f_f(\mathbf{x}) \tag{2}$$

For a *given* design point $\mathbf{x}_k$ the value of the mapping $\mathbf{z}_k = \mathbf{p}(\mathbf{x}_k)$ is obtained by solving

$$\mathbf{z}_k = \arg\min_{\mathbf{z}} |f_f(\mathbf{x}_k) - f_c(\mathbf{z})| \tag{3}$$

This minimization problem is known as the parameter extraction as it gives the set of parameters $\mathbf{z}_k$ that makes the responses of the two models match in the given design point $\mathbf{x}_k$.

As the mapping is only known from parameter extraction for one design point, $\mathbf{x}_k$, at a time it is linearized around this point in order to be able to use $f_c(\mathbf{p}(\mathbf{x}))$

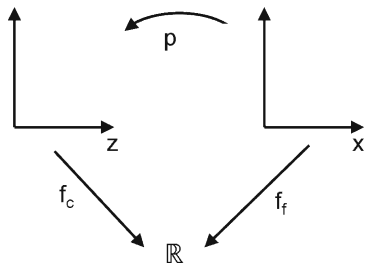

Fig. 1 Space mapping. The mapping $\mathbf{p}(\mathbf{x})$ maps the design variables such that $f_c(\mathbf{p}(\mathbf{x})) \simeq f_f(\mathbf{x})$

in the optimization problem. The linearized mapping is given by

$$\mathbf{p}_k(\mathbf{x}) = \mathbf{z}_k + \mathbf{B}_k(\mathbf{x} - \mathbf{x}_k) \tag{4}$$

where $\mathbf{B}$ is the Jacobian of $\mathbf{p}$. Due to the choice of $\mathbf{z}_k$, based on (3), the response of the surrogate model $f_c(\mathbf{p}(\mathbf{x}))$ will match that of the fine model for $\mathbf{x} = \mathbf{x}_k$.

An exact expression for the matrix $\mathbf{B}$ is not available but it is approximated using a rank one Broyden update, Broyden (1965),

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{z}_{k+1} - \mathbf{z}_k - \mathbf{B}_k\mathbf{h}_k}{\mathbf{h}_k^T\mathbf{h}_k}\mathbf{h}_k^T \tag{5}$$

where $\mathbf{h}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ is the step taken in the fine model design variables.

A similar mapping, $\mathbf{p}^{\text{con}}$, is constructed for the fine model constraint in (1). Replacing the fine model by the surrogate, the optimization problem (1) is thus replaced by a sequence of problems

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_c(\mathbf{p}_k(\mathbf{x}))) \\ \text{s.t.} \quad & c_c(\mathbf{p}_k^{\text{con}}(\mathbf{x})) = 0 \end{aligned} \tag{6}$$

Solving problem (6) yields an approximate solution, $\mathbf{x}_{k+1}$, to the real problem in (1). The mappings can then be updated through parameter extractions and subsequent updates of $\mathbf{B}$.

By choosing $\mathbf{B}_0 = \mathbf{I}$, the identity matrix, and $\mathbf{z}_0 = \mathbf{x}_0$, and similar for the constraint, the first step of the algorithm minimizes the objective function based on the coarse model, i.e. finds the optimal coarse model parameters, as $f_c(\mathbf{p}(\mathbf{x})) = f_c(\mathbf{x}_0 + \mathbf{I}(\mathbf{x} - \mathbf{x}_0))$. As the two models represent the same physical problem and we have no detailed information on how the two models correspond, this is a reasonable first step. Because the first iteration minimizes the coarse model any initial point may in principle be used. However for more complex functions local minima of the coarse model may exist.

The use of a trust region in problem (6) ensures that the linearization is only used in some desired range, i.e. $\mathbf{x}$ is only allowed to change a certain value $\frac{\delta_x}{2}$. However the trust region should not be active in the first iteration in order to allow a step to the coarse model minimizer, Søndergaard (1999). Different methods exist for updating the trust region. In Redhe and Nilsson (2006) a method which allows the region to shrink or translate depending on the design variable history is used. In the present work we use a simpler approach where the trust region is gradually increased from a small value. In the first iterations the mapping may not be very good so the design variables $\mathbf{x}$ are only allowed to change a little. As the optimization progresses and the mapping improves, the trust region is expanded. This method seems to work well on the thermal problems studied in this article. A more general and detailed description of the update of the trust region size is found in Walmag and Delhez (2005).

Depending on the functions involved several solutions to the minimization problem in the parameter extraction, (3), may exist. In order to increase the possibility for a unique solution a term including fine and coarse model gradients may be added to the residual to be minimized, Leary et al. (2001). This way also the gradients will be matched. Another way is to evaluate the fine model in a number of points $\mathbf{x}_k \pm \Delta\mathbf{x}$ close to $\mathbf{x}_k$ and add these to the residual in order to increase the possibility of a unique solution, Bandler et al. (2004). However these methods require calculation of fine model gradients or additional fine model evaluations.

The space mapping algorithm for an optimization problem with one constraint can be written:

`Do until convergence`

- Calculate new ***fine*** model design variables (from the *coarse* model with mapped design variables):

$$\begin{aligned} \mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \quad & f_c(\mathbf{p_k}^{\text{obj}}(\mathbf{x})) \\ \text{s.t.} \quad & c_c(\mathbf{p_k}^{\text{con}}(\mathbf{x})) = 0 \\ & \mathbf{x}_k - \frac{\delta_x}{2} \le \mathbf{x} \le \mathbf{x}_k + \frac{\delta_x}{2} \end{aligned}$$

- Evaluate ***fine*** model: $f_f(\mathbf{x}_{k+1})$ and $c_f(\mathbf{x}_{k+1})$
- Calculate value of mapping, $\mathbf{z}_{k+1}$, corresponding to design point $\mathbf{x}_{k+1}$ (parameter extraction):

$$\mathbf{z}_{k+1}^{\text{obj}} = \arg\min_{z^{\text{obj}}} |\mathrm{f_f}(\mathbf{x}_{\mathrm{k}+1}) - \mathrm{f_c}(\mathbf{z}^{\text{obj}})|$$

$$\mathbf{z}_{k+1}^{\text{con}} = \arg\min_{z^{\text{con}}} |\mathrm{c_f}(\mathbf{x}_{\mathrm{k}+1}) - \mathrm{c_c}(\mathbf{z}^{\text{con}})|$$

- Update B^{obj} and B^{con}
- Update trust region size δ_x

`end do`

When using space mapping for optimizing a constrained problem a total of n+2 coarse model optimization problems have to be solved for each iteration step where n is the number of constraint functions. One is the actual optimization problem using the coarse model with the mapped fine model variables. The solution of the remaining n+1 problems are parameter extractions that find the coarse model parameters that make the coarse model response match the fine model response. However all these problems are solved using the coarse model, making each solution relatively fast. From this

it is clear that not only should the coarse model be as good an approximation to the fine model as possible, it should also be fast as the number of coarse model evaluations can become very large. These properties will clearly often be contradictory. Furthermore it is advantageous if analytical derivatives can be determined for use in the optimization problem and parameter extractions.

3 Manifold mapping

Manifold mapping, Echeverria and Hemker (2006), is a recently developed method for solving computationally expensive optimization problems. As in space mapping two models are used but the mapping is performed on the coarse model response instead of on the design variables. The optimization problem to be solved is formulated as a minimization of the norm of the difference between the fine model objective and a set of specifications i.e. desired values of the objective function

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} ||\mathbf{f_f}(\mathbf{x}) - \mathbf{y}|| \tag{7}$$

where $\mathbf{x}^*$ are the optimal fine model design variables and $\mathbf{y}$ are the specifications. It should be noted that $\mathbf{f_f}$ and $\mathbf{y}$ are vectors of dimension m. A mapping $\mathbf{S}$ that approximately maps $\mathbf{f_c}(\mathbf{x}^*)$ to $\mathbf{f_f}(\mathbf{x}^*)$ is defined as the affine map

$$\mathbf{S}(\mathbf{v}) = \mathbf{f_f}(\mathbf{x}^*) + \bar{\mathbf{S}}\left((\mathbf{v}) - \mathbf{f_c}(\mathbf{x}^*)\right) \tag{8}$$

where $\mathbf{v} = \mathbf{f}_c(\mathbf{x})$, i.e., the coarse model response. Furthermore

$$\bar{\mathbf{S}} = \mathbf{J_f}(\mathbf{x}^*)\mathbf{J_c}^\dagger(\mathbf{x}^*) \tag{9}$$

where $\mathbf{J_f}(\mathbf{x}^*)$ and $\mathbf{J_c}(\mathbf{x}^*)$ are the Jacobians of the fine and the coarse model functions respectively and † denotes the pseudoinverse. This means that for $\mathbf{x} = \mathbf{x}^*$, $\mathbf{f_c}(\mathbf{x}^*)$ is mapped to $\mathbf{f_f}(\mathbf{x}^*)$ and that $\bar{\mathbf{S}}\mathbf{J_c}(\mathbf{x}^*) = \mathbf{J_f}(\mathbf{x}^*)$ i.e. the Jacobian of the coarse model is mapped to the Jacobian of the fine model. The principle of manifold mapping is shown in Fig. 2. As $\mathbf{x}^*$ is not known a priori and $\mathbf{J_f}(\mathbf{x})$ is assumed to be unavailable, the matrix $\bar{\mathbf{S}}$ can not be evaluated. Instead it is approximated at each iteration using

$$\bar{\mathbf{S}}_{k+1} = \Delta\mathbf{F_f}\Delta\mathbf{F_c}^\dagger \tag{10}$$

where $\Delta\mathbf{F_f}$ is a matrix containing changes in fine model responses during the preceding iterations and similarly

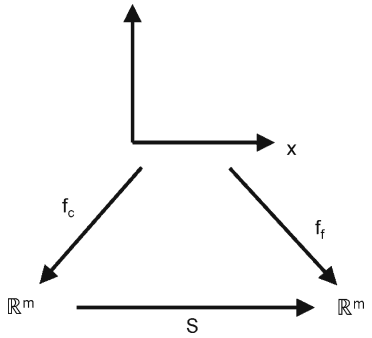

Fig. 2 Manifold mapping. In the optimum design point, $\mathbf{x}^*$, $\mathbf{S}$ maps $\mathbf{f}_c(\mathbf{x}^*)$ to $\mathbf{f}_f(\mathbf{x}^*)$ and the Jacobian $\mathbf{J_c}(\mathbf{x}^*)$ to $\mathbf{J_f}(\mathbf{x}^*)$

$\Delta\mathbf{F_c}$ contains changes in coarse model responses. For each iteration the problem

$$\begin{aligned}\mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}} \left|\left|\mathbf{S}_k(\mathbf{f_c}(\mathbf{x})) - \mathbf{y}\right|\right| \\ &= \arg\min_{\mathbf{x}} \left|\left|\Delta\mathbf{F_f}\Delta\mathbf{F_c}^\dagger(\mathbf{f_c}(\mathbf{x}) - \mathbf{f_c}(\mathbf{x}_k)) \right.\right. \\ &\qquad \left.\left. + \mathbf{f}_f(\mathbf{x}_k) - \mathbf{y}\right|\right| \end{aligned} \tag{11}$$

is solved as a substitute for problem (7). For the case of m larger than the number of design variables it can be shown, see Echeverria and Hemker (2006), that a fixed point, $\bar{\mathbf{x}}$, of the manifold mapping iteration is a local minimizer of $||\mathbf{f}_f(\mathbf{x}) - \mathbf{y}||$.

If fine model constraints are present in the optimization problem a similar mapping is created for these.

In the procedure shown above the surrogate model $\mathbf{S}_k(\mathbf{f_c}(\mathbf{x}))$ is updated at each iteration. However it is possible to rewrite the equations such that the coarse model $\mathbf{f}_c(\mathbf{x})$ is kept unchanged and the specifications $\mathbf{y}$ are updated instead. Using $\mathbf{T}_k = \bar{\mathbf{S}}_k^\dagger$ the optimization problem given by (11) is changed to

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} ||\mathbf{f_c}(\mathbf{x}) - \mathbf{y}_k|| \tag{12}$$

where $\mathbf{y}_k = \mathbf{f}_c(\mathbf{x}_k) - \mathbf{T}_k(\mathbf{f}_f(\mathbf{x}_k) - \mathbf{y})$ and $\mathbf{T}_k = \Delta\mathbf{F}_c\Delta\mathbf{F}_f^\dagger$.

Including a fine model constraint and a corresponding mapping $\mathbf{K}_k$ similar to the one given by (8) and (9) the manifold mapping algorithm can be written, Echeverria (2007):

Optimize *coarse* model:

$$\begin{aligned}\mathbf{x}_0 = \arg\min_{\mathbf{x}} &||\mathbf{f_c}(\mathbf{x}) - \mathbf{y}|| \\ \text{s.t.} \quad &\mathbf{c}_c(\mathbf{x}) = \mathbf{0}\end{aligned}$$

$\mathbf{T_0} = \mathbf{I}$
$\mathbf{K_0}(\bullet) = \mathbf{c}_f(\mathbf{x}_0) + (\bullet - \mathbf{c}_c(\mathbf{x}_0))$
`Do until convergence`

- Update specifications
  $\mathbf{y}_k = \mathbf{f_c}(\mathbf{x}_k) - \mathbf{T_k}(\mathbf{f_f}(\mathbf{x}_k) - \mathbf{y}))$
- Calculate new value of design variables $\mathbf{x}$

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} ||\mathbf{f_c}(\mathbf{x}) - \mathbf{y}_k|| \quad \text{s.t.} \quad \mathbf{K}_k(\mathbf{c_c}(\mathbf{x})) = \mathbf{0} \tag{13}$$

- Evaluate ***fine*** model and *coarse* model. Build response matrices

$$\Delta\mathbf{F_f} = \left[\mathbf{f_f}(\mathbf{x}_{k+1}) - \mathbf{f_f}(\mathbf{x}_k), ..., \mathbf{f_f}(\mathbf{x}_{k+1}) - \mathbf{f_f}(\mathbf{x}_{max(k+1-n,0)})\right]$$
$$\Delta\mathbf{F_c} = \left[\mathbf{f_c}(\mathbf{x}_{k+1}) - \mathbf{f_c}(\mathbf{x}_k), ..., \mathbf{f_c}(\mathbf{x}_{k+1}) - \mathbf{f_c}(\mathbf{x}_{max(k+1-n,0)})\right]$$
$$\Delta\mathbf{C_f} = \left[\mathbf{c}_f(\mathbf{x}_{k+1}) - \mathbf{c}_f(\mathbf{x}_k), ..., \mathbf{c}_f(\mathbf{x}_{k+1}) - \mathbf{c}_f(\mathbf{x}_{max(k+1-n,0)})\right]$$
$$\Delta\mathbf{C_c} = \left[\mathbf{c}_c(\mathbf{x}_{k+1}) - \mathbf{c}_c(\mathbf{x}_k), ..., \mathbf{c}_c(\mathbf{x}_{k+1}) - \mathbf{c}_c(\mathbf{x}_{max(k+1-n,0)})\right]$$

- Update mappings
  $\mathbf{T_{k+1}} = \Delta\mathbf{F_c}\Delta\mathbf{F_f}^{\dagger}$
  $\mathbf{K}_{k+1}(\bullet) = \mathbf{c}_f(\mathbf{x}_{k+1}) + \Delta\mathbf{C_f}\Delta\mathbf{C_c}^{\dagger}(\bullet - \mathbf{c}_c(\mathbf{x}_{k+1}))$

`end do`

It is seen that the parameter extraction step, which in space mapping requires solution of additional minimization problems, is not present in manifold mapping. Instead pseudoinverses have to be calculated which is likely to be faster, at least for problems with few degrees of freedom. This removes the risk seen in space mapping of non-unique parameter extractions, and it should be faster in most cases. This may allow the use of a more expensive coarse model as this is only used in one optimization problem. A flow diagram showing the overall procedure in space and manifold mapping is given in Fig. 3.

It is also possible to introduce a trust region, see Echeverria (2007), in the manifold mapping scheme, which will help overcome potential problems if $\Delta\mathbf{F_c}$ and $\Delta\mathbf{F_f}$ become ill-conditioned. We have not used a trust region in connection to manifold mapping in this work.

It should be noted that we did not encounter problems with infeasible constraints in this work. If the coarse model is infeasible then the optimization code used for the coarse model optimization may crash giving no new optimum. This would be problematic as there is no point to evaluate the fine model in. Some optimizers will minimize the constraint violation and thus allow the algorithm to continue (an example of this is the globally convergent version of the algorithm MMA, Svanberg (2002)). It may happen that the coarse model is feasible while the fine model is not, but as the mappings are updated during iterations to improve similarity between the two models this infeasibility should decrease.

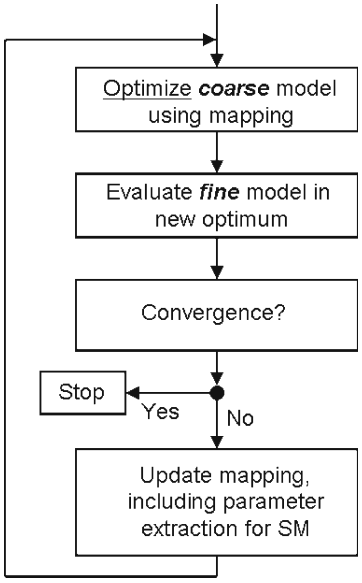

Fig. 3 Flow diagram showing the optimization loop in space and manifold mapping. For details see the text in Sections 2 and 3

4 Thermal models of FSW

As stated earlier this work focuses on the thermal aspects of FSW, more precisely the temperature field around the tool as this is of importance for the microstructure and residual stresses. Therefore, a purely thermal model can give important information about the FSW process even though the mechanical stirring is not considered in the model.

In a real welding situation there are transient phases during starting/stopping, and due to heating of the plate and machine the process may never reach complete steady state and a time-dependent solution of the heat equation should be determined. We focus here on the

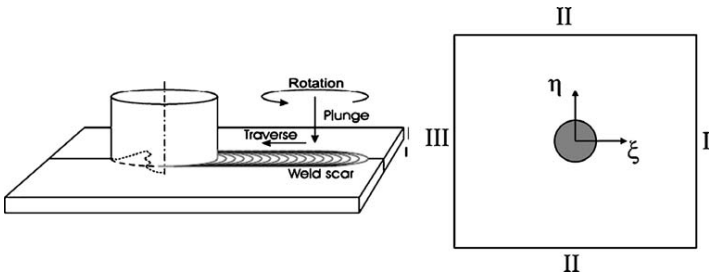

Fig. 4 *Left*: Sketch of the FSW process showing the rotating and moving tool and the workpiece, from Schmidt et al. (2004). *Right*: Sketch showing boundary conditions for the FEM model. The welding direction is towards the right. I is kept at a constant temperature, II are insulating boundaries and III is a boundary with a convective term added. The heat source is shown as a *grey circle*

phase were the tool is moving at constant velocity and is so far from the start and end points that the process can be assumed stationary. Thus a steady state model is a good approximation. A 2D finite element model has been developed in Comsol, Comsol (2006), using a simplified volume heat source as compared to the more complex heat sources in 3D models of FSW, see e.g. Schmidt and Hattel (2005c). This rather simple model also makes gradient based optimization straightforward as analytical gradients can be calculated easily. This way the results obtained using space/manifold mapping can be validated.

The stationary heat conduction equation including a convective term is written

$$\nabla(-k(T)\nabla T) = q - \rho(T)c_p(T)\mathbf{u}\nabla T \tag{14}$$

where T is the temperature, $k(T)$ is the temperature dependent conductivity, q is the heat input in W/m^3, $\rho(T)$ is the material density, $c_p(T)$ is the specific heat capacity and $\mathbf{u}$ is the welding velocity vector. The boundary conditions of the plate are chosen as indicated in Fig. 4. The right boundary, boundary I in Fig. 4, is kept at a constant temperature, while a convective term is added to the left boundary, III, such that equilibrium is obtained with the heat source q. The remaining two boundaries, II, are thermally insulating meaning there is no heat transfer across these. The model thus resembles welding in an infinitely long plate of finite width. The heat generation due to the tool is modelled as a circular heat source with the radius of the shoulder and it is assumed uniform through the plate thickness. A contourplot of the temperature field in the plate can be seen in Fig. 5.

A number of analytical solutions exist to the problem of a moving heat source on an infinite plate, Rosenthal (1946). The solution to the problem of a constant *line source* through the thickness of the plate is given by

$$T = T_0 + \frac{Q}{2\pi\, kt_{plate}} e^{-\lambda v\xi} K_0(\lambda v R) \tag{15}$$

where K_0 is the modified Bessel function of the second kind and zero'th order. T_0 is the initial plate temperature, t_{plate} is the plate thickness, $\lambda = \frac{\rho c_p}{2k}$, and R is the distance from the heat source to the point of interest while ξ is this distance projected on to the welding direction. Q is the total heat input in W.

The main difference between the FE model and the analytical model is the way the heat source is modelled. In the FE model the source is distributed over an

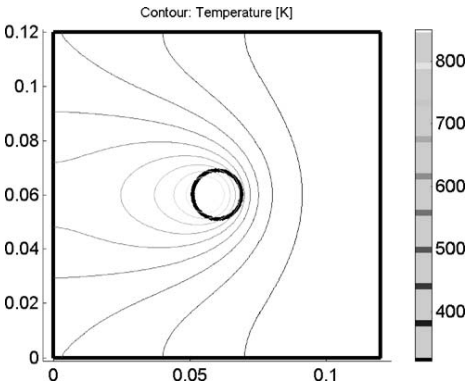

Fig. 5 Temperature field [K] around the heat source

Table 1 Material properties and plate dimensions

Property/dimension	Value
ρ	2700 $\frac{kg}{m^3}$
c_p	1000 $\frac{Ws}{kgK}$
k	167* $\frac{W}{mK}$
Side lengths	12 cm
Tool radius	9 mm
Plate thickness	3 mm

*In the examples in Section 5 the conductivity k is assumed constant whereas it is temperature dependent in the examples in Section 6

area and this leads to moderate temperature gradients around it, as seen in Fig. 5. In contrast, the analytical model assumes a line source through the thickness which results in infinite temperatures at the source. Also the boundaries in the FE model may be a cause of differences in the temperature fields, but this is most critical when points very close to the boundaries are of interest or for very low welding speeds.

In reality the material properties, e.g. k, are temperature dependent. This effect can easily be implemented in the FE model while the analytical model can only handle constant values. Thus part of the purpose of this work is to test how this influences the efficiency of space/manifold mapping when using the analytical model as the coarse model. The temperature dependence of k varies for different materials but in this work we test two different cases: constant k, Section 5, and a material with piecewise linear k as function of the temperature, Section 6. The material and model properties used in the calculations are given in Table 1.

Based on the thermal models discussed above an optimization problem is defined

$$\begin{aligned} \min_{q,v} \quad & (T(\xi,\eta) - T_1)^2 \\ \text{s.t.} \quad & T_{Avg} = T_2 \end{aligned} \tag{16}$$

In order to control the temperature field we want to minimize the difference between the temperature in a given point (ξ,η) and a specified temperature T_1. This is a way of controlling the size of the heat affected zone (HAZ). Note that we are not trying to minimize the size of the HAZ. For detailed studies of microstructure evolution the temperature gradient may play an important role and this should be reflected in the choice of objective function. In this work we choose $T_1 = 300°C$ in order to control the size of the HAZ. However, for practical applications, this temperature should be chosen based on the actual welding situation and the material being welded. The constraint is a specified average temperature under the tool which can be used as a rough measure of a successful weld, Tutum et al. (2007). In this work $T_2 = 500°C$, which is chosen based on experiments that show that – as a rule of thumb – one obtains a good trade-off between no melting and softness of the material (which makes the stirring together of the two materials possible). The average temperature is easily calculated in Comsol using a built-in integration function. For the analytical coarse model the average temperature is found by summation of temperatures in points under the tool. The design variables are the total heat input q and the welding speed v.

5 Example: influence of model size

In this section we show tests on how the FE model size, i.e. number of degrees of freedom in the fine model, influences the computation time for space/manifold mapping and gradient based optimization using the full model. All optimization problems in this paper are solved using SNOPT, Gill et al. (2006). For the gradient based method using the full fine model, analytical sensitivities were calculated using the direct method Choi and Kim (2005). Two different coarse models are used in the space mapping; one is the analytical model given in (15) and the other is a FE model with only 5032 degrees of freedom. The fine models tested have between 13897 and 487812 degrees of freedom. The problem given by (16) is solved with $\xi = 0\,m$, $\eta = 0.005\,m + R_{tool} = 0.014\,m$, $T_1 = 573\,K$ and $T_2 = 773\,K$. The optima for the two different coarse models and for the fine models are shown in Table 2 for reference.

As expected we see that the optimum for the analytical model is quite far from that of the low density mesh FE model. The optimum of this model is on the other hand close to that of the high density mesh FE models. We note that the optimum changes slightly as the mesh density is increased.

Table 2 Optima for the different models used in the examples in Section 5

		q^*[W]	v^* [$\frac{mm}{s}$]
Coarse	Analytical	684.4	2.91
Coarse	FEM 5032 dofs	1064.5	5.12
Fine	FEM 13897 dofs	1022.9	4.94
Fine	FEM 31227 dofs	1012.3	4.90
Fine	FEM 55410 dofs	1006.9	4.87
Fine	FEM 124791 dofs	1003.4	4.86
Fine	FEM 497812 dofs	1001.4	4.85

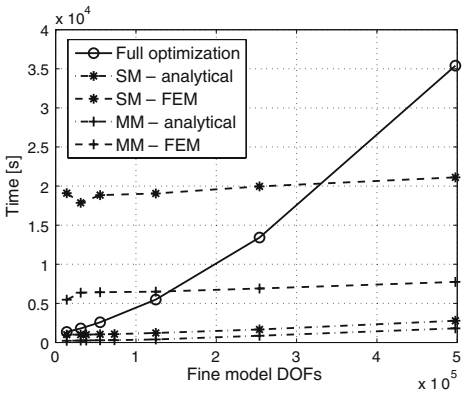

Fig. 6 Calculation time for different optimization methods. The starting point for the optimization is $q = 9000\,W$ and $v = 4\,\frac{mm}{s}$. Full optimization indicates gradient based optimization directly using the fine model

The first step in the space/manifold mapping will lead to the coarse model optimum meaning that any starting point may be used. In this example the conductivity k is assumed constant. The solution time for different model sizes and the different optimization methods are shown in Fig. 6.

Looking at Fig. 6 it is seen that the time spent by the gradient based method using the full fine model increases rapidly as the model size increases. The increase is much faster than for space/manifold mapping. This is due to the fact that also gradients are calculated and thus two more solutions of the equation system are required, compared to only one function evaluation needed for space/manifold mapping.[2] In the examples shown above it was found that using a coarse-mesh FE model as coarse model in space/manifold mapping requires fewer iterations than using an analytical model but as the coarse-mesh model in itself is much more expensive than the analytical model, the total CPU-time is larger. This affects not only the coarse model optimization but also the parameter extraction steps in space mapping. Thus the difference between space and manifold mapping becomes very large because the parameter extraction is not present in manifold mapping. Table 3 shows the number of iterations, equaling the number of fine model evaluations, required.

[2]We note that using an adjoint method would not change this.

Table 3 Number of iterations required for space/manifold mapping and two different coarse models

Space mapping		Manifold mapping	
Analytical	FEM	Analytical	FEM
9	6	6	3

The number of fine model evaluations equals the number of iterations

Looking at the plots in Fig. 7 it is seen that the difference between space and manifold mapping seems to be constant for all fine model sizes. By closer inspection it was found that the time per iteration converges to the same value in the case of an analytical coarse model. This means that the fine model evaluation becomes the dominant time consumer and greatly overshadows the

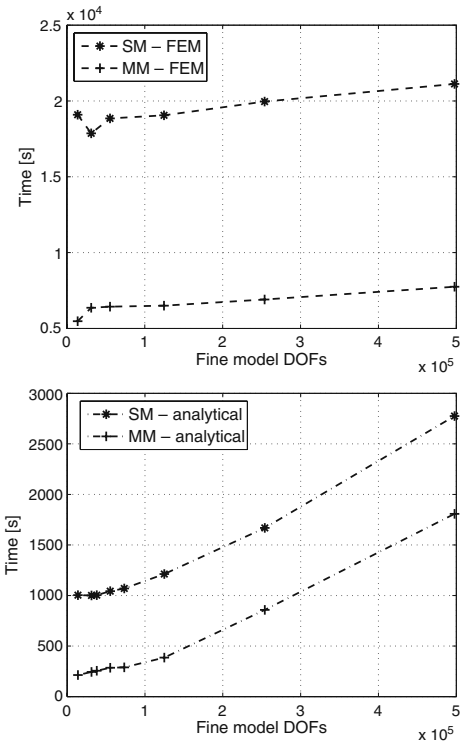

Fig. 7 Zoom of Fig. 6. *Top*: results using the FEM model as coarse model. *Bottom*: results using the analytical coarse model

time spent on parameter extraction. This effect was not as obvious in the case of a FE coarse model because the optimization of the coarse model and the parameter extractions are responsible for a large part of the total time spent, even for large fine models.

As stated earlier the average temperature in the analytical model is calculated by summation of temperatures in a number of points (276 points for the examples shown above). Thus varying the number of points will influence the quality of the coarse model and possibly influence the optimization time. In order to study this in more detail a number of space mapping calculations were performed using between 1184 and 4 points for calculation of the average temperature. In all cases except the one with only four points nine iterations were required. The last case required 10 iterations meaning that even a relatively small number of points will be sufficient.

It should be noted that the parameter extraction steps in space mapping, where the coarse model parameters are determined, are performed using the actual temperatures $T(\xi, \eta)$, based on the fine and coarse models, instead of the objective function values f_f and f_c. This is due to the formulation of the objective function, i.e. the squared temperature difference, which can take the same value for two different temperatures. It is therefore better to match the temperatures themselves.

Due to the fact that the optimal solution satisfies the constraint and the objective function takes the value 0, the problem could also have been solved using another formulation in manifold mapping. The temperature in the point and the average temperature could both be included in the responses $\mathbf{f}_f$ and $\mathbf{f}_c$ while setting $\mathbf{y} = [T_1 \quad T_2]^T$. Including the constraint in the objective will not make sense in all cases however.

In the examples above, two different coarse models have been used; an analytical and a coarse-mesh FE model. As mentioned earlier an alternative is to use a response surface, as done in Lahaye et al. (2007), Jansson et al. (2003), based on e.g. kriging, radial basis functions or polynomials. It will be especially attractive for very complex fine models where analytical approximations do not exist or coarse FE models become too expensive, or because analytical gradients cannot be calculated or be relied on. An interesting feature in this context is that the response surface can be updated almost for free using the fine model evaluation performed in each iteration of space/manifold mapping. The optimization problem treated in this paper was also solved using a radial basis function response surface as coarse model in combination with space/manifold mapping. However the advantages of this approach come to their right when dealing with more complex models.

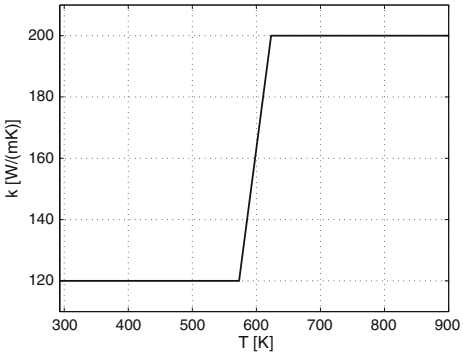

Fig. 8 Conductivity as function of temperature

6 Example: Temperature dependent conductivity

In this section the problem complexity is increased by using a temperature dependent conductivity $k(T)$ while the remaining material properties, ρ and c_p, are kept constant. This is only possible in the FE model so the analytical model will be a worse approximation than in the previous examples. The temperature dependence is modelled using a piecewise linear function. This is a simplification of a real material but it will serve as an example of a more complicated material behaviour. For a description of the actual temperature dependent behaviour of different materials see e.g. Olafsson et al, (1997). Figure 8 shows how the conductivity depends on the temperature.

To test if the analytical coarse model will still be useful, the optimization problem is solved using the temperature dependent conductivity. To further test the flexibility of the two techniques the constant conductivity in the coarse model is varied. The results in form of number of fine model evaluations required to reach convergence are shown in Table 4. The values $k = 120\,\frac{W}{mK}$, $k = 200\,\frac{W}{mK}$ and $k = 167\,\frac{W}{mK}$ were chosen

Table 4 Results for temperature dependent conductivity

$k\,[W/(mK)]$	Fine model evaluations	
	Space mapping	Manifold mapping
120	16	13
167	11	9
200	10	8

k indicates the value used in the analytical coarse model

as the extreme values of the temperature dependent conductivity and as the value used in the previous example, respectively.

As seen in the previous examples, manifold mapping requires fewer iterations than space mapping. The physics represented by the problem using the low conductivity makes it harder to solve the case where this is used as the coarse model (the fine FE model contains large regions with a k that is not 120). In Fig. 5 the temperature field for constant k is shown. Using a temperature dependent conductivity will change this to some extend but the overall shape of the isotherms will remain the same. Based on the given optimization problem the temperature under the tool should be higher than 573 K where the conductivity increases abruptly. This means that in the area where the average temperature is calculated the conductivity is far from the low value used in the coarse model. Contrary, the use of a higher conductivity in the coarse model will increase the similarity of the two models and this is reflected in the required number of iterations.

7 Conclusion

The use of space and manifold mapping for optimization of a thermal FE model of a FSW process has been studied. Two different coarse models were used; an analytical model and a coarse-mesh FE model. It was found that using the coarse-mesh FE model required fewer iterations and thereby fewer fine model evaluations than when using the analytical model as coarse model. This can be explained by the higher degree of similarity between the two FE models. However the coarse-mesh FE model was so computationally expensive that the use of the analytical model was faster despite the increased number of iterations. For fine FE models with a large number of degrees of freedom the use of space/manifold mapping proved to be significantly faster than gradient based optimization using the full model. It was found that manifold mapping generally required fewer iterations, and thus fine model evaluations, than space mapping. This contributed to making manifold mapping faster than space mapping along with the fact that parameter extractions are not required.

A more challenging problem in which temperature dependent material data was used in the fine model was also solved. Using the analytical model as coarse model it was found that this example required more iterations than the examples with constant material data.

The main advantage of space/manifold mapping is the fact that fewer fine model evaluations are required and that no gradient information regarding the fine model is used. Thus the use of space/manifold mapping appears promising for optimization of more complex models of the FSW process provided that effective coarse models are available. In case analytical or coarse FE models are not applicable, response surfaces may be used as coarse models.

This is to our knowledge the first attempt to use continuous optimization techniques to optimize process parameters for FSW. The work should thus be seen as a feasibility study. Future work will include matching and validation with experimental tests. Also, a major issue is the robustness of the obtained parameters, and here it would be natural to implement ideas from robust optimization. A study of how this can be implemented in the current framework of space and manifold mapping is currently in progress.

Acknowledgements This work is part of the InnoJoint project sponsored by the Danish Research Council for Technology and Production Sciences, contract number 274-05-0232.

References

Bakr M, Bandler J, Madsen K, Søndergaard J (2001) An introduction to the space mapping technique. Optim Eng 2: 369–384

Bandler J, Biernacki R, Chen S, Grobelny P, Hemmers R (1994) Space mapping technique for electromagnetic optimization. IEEE Trans Microwave Theor Tech 42(12):2536–2544

Bandler J, Cheng Q, Dakroury S, Mohamed A, Bakr M, Madsen K, Søndergaard J (2004) Space mapping: the state of the art. IEEE Trans Microwave Theor Tech 52(1):337–361

Broyden C (1965) A class of methods for solving nonlinear simultaneous equations. Math Comput 19(92):577–593

Chen C, Kovacevic R (2003) Finite element modeling of friction stir welding—thermal and thermomechanical analysis. Int J Mach Tools Manuf 43:1319–1326

Choi K, Kim NH (2005) Structural sensitivity analysis and optimization, vol 1 & 2. Springer, New York

Colegrove P, Shercliff H, Threadgill P (2003) Modelling and development of the trivex(tm) friction stir welding tool. In: 4th International Symposium on Friction Stir Welding, Park City, 14–16 May 2003

Comsol (2006) Comsol website. www.comsol.com

Echeverria D (2007) Two new variants of the manifold-mapping technique. COMPEL 26(2):334–344

Echeverria D, Hemker P (2005) Space mapping and defect correction. Comput Methods Appl Math 5(2):107–136

Echeverria D, Hemker P (2006) On the manifold-mapping optimization technique. Technical report mas-r0612, Centrum voor Wiskunde en Informatica (CWI), MAS

Gill P, Murray W, Saunders M (2006) User's guide for snopt version 7: software for large-scale nonlinear programming

Jansson T, Nilsson L, Redhe M (2003) Using surrogate models and response surfaces in structural optimization - with application to crashworthiness design and sheet metal forming. Struct Multidisc Optim 25:129–140

Khandkar M, Khan J (2001) Thermal modeling of overlap friction stir welding for al-alloys. J Mater Process Manuf Sci 10:91–105

Lahaye D, Canova A, Grusso G, Repetto M (2007) Adaptive manifold-mapping using multiquadric interpolation applied to linear actuator design. COMPEL 26(2):225–235

Leary S, Bhaskar A, Keane A (2001) A constraint mapping approach to the structural optimization of an expensive model using surrogates. Optim Eng 2:385–398

Olafsson P, Sanström R, Karlsson A (1997) Comparison of experimental, calculated and observed values for electrical and thermal conductivity of aluminium alloys. J Mater Sci 32:4383–4390

Redhe M, Nilsson L (2006) A multipoint version of space mapping optimization applied to vehicle crashworthiness design. Struct Multidisc Optim 31:134–146

Rosenthal D (1946) The theory of moving sources of heat and its application to metal treatments. Transactions of the ASME, November

Schmidt H, Hattel J (2005a) Cfd modelling of the shear layer around the tool probe in friction stir welding. Friction Stir Welding and Processing III, TMS, San Francisco, 13–17 February 2005

Schmidt H, Hattel J (2005b) A local model for the thermomechanical conditions in friction stir welding. J Mater Process Manuf Sci 13:77–93

Schmidt H, Hattel J (2005c) Modelling heat flow around tool probe in friction stir welding. Sci Technol Weld Join 10(4):176–186

Schmidt H, Hattel J, Wert J (2004) An analytical model for the heat generation in friction stir welding. Model Simul Mater Sci Eng 12:143–157

Søndergaard J (1999) Non-linear optimization using space mapping. Master's thesis, IMM, Technical University of Denmark

Svanberg K (2002) A class of globally convergent optimization methods based on conservative convex separable approximations. SIAM J Optim 12(2):555–573

Tutum CC, Schmidt H, Hattel J, Bendsøe MP (2007) Estimation of the welding speed and heat input in friction stir welding using thermal models and optimization. In: Kwak BY et al. (eds) Proc 7th world congress on structural and multidisciplinary optimization. COEX Seoul, Korea, pp 2639–2646, 21–25 May 2007

Walmag J, Delhez E (2005) A note on trust-region radius update. SIAM J Optim 16(2):548–562

Appendix B

Article 2

Anders Larsen; Domenico Lahaye; Henrik Schmidt; Jesper Hattel; Martin Bendsøe. Robust Optimization of Thermal Aspects of Friction Stir Welding Using Manifold Mapping Techniques *12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, British Columbia, Sep. 10-12, 2008*, AIAA-2008-6004

Robust Optimization of Thermal Aspects of Friction Stir Welding Using Manifold Mapping Techniques

Anders A. Larsen*
Dept. of Mathematics, Tech. University of Denmark, Lyngby Denmark
Domenico Lahaye
Delft Institute for Applied Mathematics, Tech. University of Delft, Netherlands
Henrik Schmidt and Jesper Hattel
Dept. of Mechanical Engineering, Tech. University of Denmark, Lyngby Denmark
Martin P. Bendsøe
Dept. of Mathematics, Tech. University of Denmark, Lyngby Denmark

The aim of this paper is to optimize a friction stir welding process taking robustness into account. The optimization problems are formulated with the goal of obtaining desired mean responses while reducing the variance of the response. We restrict ourselves to a thermal model of the process and use the manifold mapping technique to solve the optimization problems using a fast analytical coarse and an expensive accurate fine model. The statistics of the response are calculated using Taylor expansions and are compared to Monte Carlo simulations. The results show that the use of manifold mapping reduces the number of fine model evaluations required and that the Taylor expansion approach gives good results when compared to Monte Carlo simulations.

Nomenclature

$\mathbf{b}$	Vector of uncertain parameters
c_p	Specific heat capacity, $\frac{J}{kgK}$
$\Delta\mathbf{F_f}$	Fine model response matrix
$\Delta\mathbf{F_c}$	Coarse model response matrix
$\mathbf{f_f}$	Fine model response
$\mathbf{f_c}$	Coarse model response
$\mathbf{J_f}$	Fine model Jacobian
$\mathbf{J_c}$	Coarse model Jacobian
K_0	Bessel function of the second kind and zeroth order
k	Thermal conductivity, $\frac{W}{mK}$
q, Q	Heat source, $\frac{W}{m^3}$, W
R	Distance from heat source to point, m
R_λ, r_λ	Stabilization parameters
$\mathbf{S}$, $\mathbf{T}$	Mapping
T	Temperature, K
t_{plate}	Plate thickness, m
$\mathbf{U}_{\Delta\mathbf{F_f}}, \mathbf{U}_{\Delta\mathbf{F_c}}, \mathbf{V}$	Generalized singular value decomposition matrices
$\mathbf{u}$	Welding velocity vector, $\frac{m}{s}$
v	Welding speed in x-direction, $\frac{m}{s}$
w	Weight factor
$\mathbf{x}$	Vector of design variables

*Email: a.a.larsen@mat.dtu.dk

$\mathbf{y}$	Specifications of optimization problem
α, δ, τ	Stabilization parameters
λ	Thermal property $\frac{\rho c_p}{2k}$ and stabilization parameter
μ	Mean value
ρ	Material density, $\frac{kg}{m^3}$
$\mathbf{\Sigma}_{\Delta\mathbf{F_f}}, \mathbf{\Sigma}_{\Delta\mathbf{F_e}}$	Generalized singular value decomposition matrices
σ	Standard deviation, singular values
ξ, η	Plate coordinates, m
Subscript	
i	Parameter number
j	Parameter number
k	Iteration number

I. Introduction

Friction stir welding (FSW) is a solid state welding process and was invented in the early 1990's by The Welding Institute[30]. A rotating tool is moved along the weldline between the two parts to be welded creating heat and causing softening of the material such that the two parts are joined due to large plastic deformations of the material, see Figure 1. Temperatures are generally well below the melting temperature but localized melting may occur close to the tool. The low temperatures are one of the advantages of FSW compared to conventional welding methods.

A number of different models are used for modelling FSW, e.g. thermal models[13,28], CFD models[6,26] and thermomechanical models[4,27]. This paper only focuses on the thermal aspects of FSW by investigating the temperature field around a heat source resembling the heat created at the tool. The temperature field is of importance for e.g. residual stresses and microstructure development.

Robust optimization, here meaning optimization with the purpose of obtaining a desired mean while reducing the variance of the response, is a challenging problem due to the need for calculation of mean and variance of responses. An overview of robust optimization is given in Ref. 3 where different approaches to robust optimization are discussed. The paper also describes a widely used approximation for calculating the statistics of the response based on Taylor expansions. This method is applied in the present paper. Ref. 23 also reviews different strategies for robust optimization and discusses different formulations for the optimization problem. In Ref. 14 the Taylor expansion method is discussed in connection to six sigma design. Examples concerning crashworthiness of a car are shown. Ref. 17 uses a different approach for obtaining mean and variance by first constructing a kriging surrogate model of responses. This surrogate model is then used for creating a new kriging model of variance based on Monte Carlo simulations. The advantage compared to the Taylor expansion approach is that no gradients are required for the estimation of mean and variance. Ref. 22 uses a sigma point method where mean and variance are estimated using function values only in a number of points around the design point of interest. This way no gradients are required for the evaluation of mean and variance and the accuracy of the mean value is reported to be higher than when using a first order Taylor expansion method. In Ref. 24 the author describes methods for feasibility robustness, i.e., robustness with respect to constraint satisfaction, and sensitivity robustness where the sensitivity of the design with respect to variations is minimized. First order Taylor expansions are used for calculating mean and variance.

In this work we use the manifold mapping technique[10] for solving the optimization problems. The idea central to manifold mapping is to combine an accurate and an approximate model for the optimization. A mapping is established between the the space of responses of the two models in such a way that most of the computational work can be performed on the coarse model. In this way computation time is reduced. Another advantage is that the user is not required to provide gradient information of objective and constraint functions as this information is iteratively approximated during iterations. This derivative-free property is interesting in the present context as the objective and constraint functions include statistical data that are based on Taylor expansions and thereby derivatives of response functions. The use of Comsol Multiphysics[7] allows sensitivities based on analytical expressions to be computed efficiently. In Ref. 15 manifold mapping is successfully applied to optimization of a linear actuator using a response surface as coarse model. Ref. 9 describes a manifold mapping version using a stabilization method to improve convergence for strongly

nonlinear problems and a version that solves constrained optimization problems.

The numerical examples included in the paper are intended to show the effects of including uncertainties in the optimization problem as well as the usefullness of manifold mapping for this type of problems. We consider problems with one and three uncertain parameters and compare the results to Monte Carlo simulations and use two different formulations of the optimization problem. The efficiency of the manifold mapping technique is shown by a small study comparing the number of fine model evaluations required using manifold mapping and more traditional optimization methods.

The article is built up as follows. Section II describes the thermal models of the FSW process. Section III focuses on robust optimization and the evaluation of mean and variance of the response followed by Section IV that describes the manifold mapping technique used in the paper as well as the formulations of the optimization problems. Numerical examples are shown in Sections V, VI and VII. Section VIII deals with the computational efficiency of manifold mapping compared to direct methods while Section IX concludes the paper.

II. Friction stir welding

The present work focuses on optimization of a 2D thermal model of the FSW process not taking e.g. mechanical effects into account. This is a simplification of a true welding situation, but the temperature field is of great importance to e.g. residual stresses and microstructure. For regions of the plate that are not too close to the tool a pure thermal model may be sufficient to model microstructural changes. Also the simplified model reduces computation times thus allowing more detailed studies using Monte Carlo methods for validation of results. The stationary heat conduction equation including a convective term is written

$$\nabla(-k\nabla T) = q - \rho\, c_p \mathbf{u} \cdot \nabla T \tag{1}$$

where T is the temperature, k is the thermal conductivity, q is the source term in W/m^3, ρ is the material density, c_p is the specific heat capacity and $\mathbf{u} = [v \quad 0]^T$ is the welding velocity vector. q and v are the design variables used throughout the article. The boundary conditions of the plate are chosen as indicated in Fig. 1. The right boundary (I) is kept at a constant temperature, while a convective term is added to the left boundary (III) such that equilibrium is obtained with the heat source q. The remaining two boundaries (II) are thermally insulating meaning there is no heat transfer across these. The model thus resembles welding in an infinitely long plate of finite width. The heat generation due to the tool is modelled as a circular heat source with the radius of the tool and it is assumed uniform through the plate thickness. This model is implemented in Comsol. A number of analytical solutions exist to the problem of a moving heat source on

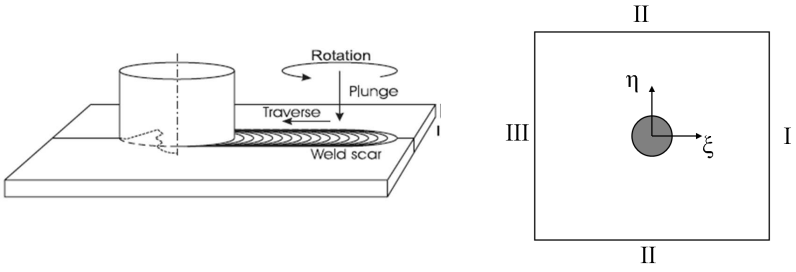

Figure 1. Left: Sketch of the FSW process showing the rotation and translation of the tool as well as the workpiece, from Ref. 27. Right: Sketch showing boundary conditions for the FEM model. The welding direction is towards the right. I is kept at a constant temperature, II are insulating boundaries and III is a boundary with a convective term added. The heat source is shown as a grey circle.

an infinite plate[25]. The solution to the problem of a constant *line source* through the thickness of the plate

is given by

$$T = T_0 + \frac{Q}{2\pi\, k\, t_{plate}} e^{-\lambda v \xi} K_0(\lambda v R) \tag{2}$$

where K_0 is the modified Bessel function of the second kind and zeroth order. T_0 is the initial plate temperature, t_{plate} is the plate thickness, $\lambda = \frac{\rho c_p}{2k}$, and $R = \sqrt{\xi^2 + \eta^2}$ is the distance from the heat source to the point of interest while ξ is the welding direction component of R. Q is the total heat input in W and v is the welding speed. The main differences between the analytical and the numerical model are the boundary conditions and the modelling of the heat source. The analytical model assumes an infinitely large plate whereas the numerical model has thermally insulating boundaries parallel to the welding direction. As seen in Figure 1 the heat source is modelled as a distributed area source in contrast to the analytical that model assumes a line source through the thickness of the plate leading to infinite temperatures at the source. Another very important difference in this context is the computation time of the two models. Using manifold mapping a large number of coarse model, i.e., the Rosenthal solution, evaluations are required but due to the simplicity of the model this is a computationally inexpensive task. An alternative to the analytical model is to use a response surface, based on a number of evaluations of an FE model[15] or a low mesh density FE model as coarse model.

III. Robust optimization

The purpose of robust optimization as presented here is to obtain a desired mean response and at the same time to reduce the variance of the response due to the effect of uncertain parameters. We only focus on the case where the variance of the uncertain parameters cannot be changed. In some cases the variance of uncertain parameters may be controlled through e.g. more advanced (and, often, costly) manufacturing processes or control of machinery.

Robust designs can be achieved in different ways but a widely used approach is to minimize a combination of mean, μ_f, and standard deviation, σ_f, of some scalar function f[3]. Thus different formulations are possible, e.g.

$$\min\ w\mu_f + (1-w)\sigma_f \quad , \quad 0 \le w \le 1 \tag{3}$$

where the weighting factor w is chosen by the designer. The value of w could be based on experience or alternatively a Pareto front may be created to assist the choice. The function f could for instance be given by an expression like $f(T(\mathbf{b})) = (T(\mathbf{b}) - T_1)^2$ where T is a response and T_1 is a desired value of that response. $\mathbf{b}$ are the uncertain parameters which may or may not include design variables.

Another possible formulation of the optimization problem is

$$\begin{aligned} \min \quad & \sigma_f^2 \\ \text{s.t.} \quad & \mu_f \le \mu_0 \end{aligned} \tag{4}$$

In this formulation the weighting w is not present but instead the constant μ_0 that limits the allowable mean has to be determined by the designer.

One of the major challenges in robust optimization is the evaluation of mean and standard deviation of the response. These values are given by[21]

$$\begin{aligned} \mu_f &= \int_{-\infty}^{\infty} f(\mathbf{b})\, p_b(\mathbf{b})\, d\mathbf{b} \\ \sigma_f^2 &= \int_{-\infty}^{\infty} (f(\mathbf{b}) - \mu_f)^2\, p_b(\mathbf{b})\, d\mathbf{b} \end{aligned} \tag{5}$$

where p_b is the joint probability density function. For statistical independent variables $p_b = p_{b_1} \cdot p_{b_2} \cdots p_{b_n}$,[23] i.e., the product of the individual probability density functions. Using Eq. (5) directly is often not possible for practical problems so different approximations are applied. An often used approach is based on Taylor expansions of first or second order. The mean and standard deviation, for statistical independent variables, can be calculated from[14]

$$\begin{aligned} \mu_f &= f(\mu_{\mathbf{b}}) \\ \sigma_f^2 &= \sum_{i=1}^{n} \left(\frac{\partial f}{\partial b_i}(\mu_{\mathbf{b}}) \right)^2 \sigma_{b_i}^2 \end{aligned} \tag{6}$$

for a first order expansion and from

$$\begin{aligned}\mu_f &= f(\mu_{\mathbf{b}}) + \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2 f}{\partial b_i^2}(\mu_{\mathbf{b}})\,\sigma_{b_i}^2 \\ \sigma_f^2 &= \sum_{i=1}^{n}\left(\frac{\partial f}{\partial b_i}(\mu_{\mathbf{b}})\right)^2\sigma_{b_i}^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial^2 f}{\partial b_i \partial b_j}(\mu_{\mathbf{b}})\right)^2\sigma_{b_i}^2\sigma_{b_j}^2\end{aligned} \tag{7}$$

for a second order expansion. μ_{b_i} and σ_{b_i} are mean and standard deviation of the uncertain parameter b_i. From Eq. (6) it is clear that if the function $f(\mathbf{b})$ has a stationary point, i.e., $\frac{\partial f}{\partial b_i} = 0\,\forall\, i$, then the calculated standard deviation will be zero. This is only a correct evaluation if f is in fact a constant, and is clearly erroneous in general.

The use of Taylor expansions allows us to evaluate the statistical data to be used in the robust optimization. Using Comsol the derivatives can be calculated efficiently at the cost of additional solutions of the equation system. We note that it may not always be possible to calculate reliable derivatives efficiently and that the method may not be appropriate for noisy models. A more direct method for evaluating the mean and standard deviation is through Monte Carlo simulations. For large sample sizes this method yields accurate measures of mean and standard deviation but the computational cost is correspondingly high, especially when compared to the Taylor expansion approach. However, different sampling strategies may be used to reduce the number of evaluations required for a trustworthy result. One such method is Latin Hypercube sampling(LHS)[12,20] which is reported to require an order of magnitude fewer points than plain Monte Carlo simulations for similar accuracy. The idea is to divide the individual probability distributions into a number of subsets with equal probability. The sample points are selected such that each subset of each variable is only used once. The software package DACE[18] is used for creating the LHS sample points used in the Monte Carlo simulations.

IV. Manifold mapping

Manifold mapping[11] is a recently developed method for solving computationally expensive optimization problems. The method is related to the space mapping technique[1,2] in the sense that it uses two models for the optimization. One is the actual model to be optimized, called the *fine model*, that is often computationally expensive to evaluate and for which the computation of gradients for use in classical optimization methods might be impossible. The other model used, the *coarse model*, is faster to evaluate but less accurate. They both model the same physical problem and have the same number of design variables and responses. In this article we use as fine and coarse model the FE and the Rosenthal models described in Section II, respectively. A mapping $\mathbf{S}$ is defined such that the response of the coarse model, $\mathbf{f_c}$, at the optimum design point, $\mathbf{x}^*$, is mapped to the corresponding response of the fine model, $\mathbf{f_f}$, i.e., $\mathbf{S}(\mathbf{f_c}(\mathbf{x}^*)) = \mathbf{f_f}(\mathbf{x}^*)$. This is in contrast to classical space mapping where the mapping is performed on design variables rather than responses. As $\mathbf{S}$ maps the coarse model response to the fine model response, the mapped coarse model can be used as a surrogate for the fine model when solving the optimization problems. In this work we use two different variants of the manifold mapping technique, a stabilized variant that controls the stepsize in the optimization process and a variant that allows optimization of problems with expensive constraints.

The optimization problem to be solved is formulated as a minimization of the norm of the difference between the fine model response and a set of specifications, $\mathbf{y}$, i.e., desired values of the response function

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}}\,||\mathbf{f_f}(\mathbf{x}) - \mathbf{y}||. \tag{8}$$

It should be noted that $\mathbf{f_f}$ and $\mathbf{y}$ are vectors of dimension m. The mapping $\mathbf{S}$ that maps $\mathbf{f_c}(\mathbf{x}^*)$ to $\mathbf{f_f}(\mathbf{x}^*)$ is defined as the affine map

$$\mathbf{S}(\mathbf{v}) = \mathbf{f_f}(\mathbf{x}^*) + \bar{\mathbf{S}}(\mathbf{v} - \mathbf{f_c}(\mathbf{x}^*)) \tag{9}$$

where $\mathbf{v} = \mathbf{f_c}(\mathbf{x})$, i.e., the coarse model response. Furthermore

$$\bar{\mathbf{S}} = \mathbf{J_f}(\mathbf{x}^*)\mathbf{J_c}^{\dagger}(\mathbf{x}^*) \tag{10}$$

where $\mathbf{J_f}(\mathbf{x}^*)$ and $\mathbf{J_c}(\mathbf{x}^*)$ are the Jacobians of the fine and the coarse model functions respectively and † denotes the pseudoinverse. This means that for $\mathbf{x} = \mathbf{x}^*$, $\mathbf{f_c}(\mathbf{x}^*)$ is mapped to $\mathbf{f_f}(\mathbf{x}^*)$ and that $\bar{\mathbf{S}}\mathbf{J_c}(\mathbf{x}^*) =$

$\mathbf{J_f}(\mathbf{x}^*)$, i.e., the Jacobian of the coarse model is mapped to the Jacobian of the fine model. As $\mathbf{x}^*$ is not known a priori and $\mathbf{J_f}(\mathbf{x})$ is assumed to be unavailable, the matrix $\bar{\mathbf{S}}$ can not be evaluated. Instead it is approximated at each iteration using

$$\bar{\mathbf{S}}_k = \Delta\mathbf{F_f}_k \Delta\mathbf{F_c}_k^{\dagger} \tag{11}$$

where $\Delta\mathbf{F_f}$ is a matrix containing changes in fine model responses during the preceding iterations and similarly $\Delta\mathbf{F_c}$ contains changes in coarse model responses. For each iteration the problem

$$\begin{aligned} \mathbf{x}_{k+1} &= \underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{S}_k(\mathbf{f_c}(\mathbf{x})) - \mathbf{y}|| \\ &= \underset{\mathbf{x}}{\operatorname{argmin}} ||\Delta\mathbf{F_f}_k \Delta\mathbf{F_c}_k^{\dagger}(\mathbf{f_c}(\mathbf{x}) - \mathbf{f_c}(\mathbf{x}_k)) \\ &\qquad + \mathbf{f_f}(\mathbf{x}_k) - \mathbf{y}|| \end{aligned} \tag{12}$$

is solved as a substitute for problem (8). This is an optimization problem involving only the coarse model and the fine model response from the previous iteration.

Following the procedure outlined above the surrogate model $\mathbf{S}_k(\mathbf{f_c}(\mathbf{x}))$ is updated at each iteration. An alternative procedure can be obtained by rewriting the equations such that the specifications $\mathbf{y}$ are updated and the coarse model is kept unchanged. By using $\mathbf{T}_k = \bar{\mathbf{S}}_k^{\dagger}$ the updated specifications are denoted $\mathbf{y}_k$ and the optimization problem given in eq. (12) is changed to

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \parallel \mathbf{f_c}(\mathbf{x}) - \mathbf{y}_\mathrm{k} \parallel \tag{13}$$

where $\mathbf{y}_k = \mathbf{f_c}(\mathbf{x}_k) - \mathbf{T}_k(\mathbf{f_f}(\mathbf{x}_k) - \mathbf{y})$ and $\mathbf{T}_k = \Delta\mathbf{F_c}_k \Delta\mathbf{F_f}_k^{\dagger}$.

To overcome problems with badly conditioned matrices a Levenberg-Marquard type stabilization may be included in the method[9] to control the steps taken in the design variables between iterations. The manifold mapping algorithm including stabilization can be written:
Optimize the *coarse* model:

$$\mathbf{x}_0 = \underset{\mathbf{x}}{\operatorname{argmin}} \parallel \mathbf{f_c}(\mathbf{x}) - \mathbf{y} \parallel$$

$\mathbf{T_0} = \mathbf{I}$
$\lambda_0 = 1$
`Do until convergence`

- Update specifications

$$\mathbf{y}_k = \mathbf{f_c}(\mathbf{x}_k) - \mathbf{T}_k(\mathbf{f_f}(\mathbf{x}_k) - \mathbf{y}) \frac{1}{1 + \delta\lambda_k} \tag{14}$$

- Calculate new value of design variables $\mathbf{x}$

$$\mathbf{x}_{k+1} \quad = \quad \underset{\mathbf{x}}{\operatorname{argmin}} \parallel \mathbf{f_c}(\mathbf{x}) - \mathbf{y}_\mathrm{k} \parallel$$

- Only accept steps that decrease objective function
  `while` $\parallel \mathbf{f}_f(\mathbf{x}_{k+1}) - \mathbf{y} \parallel > \alpha \parallel \mathbf{f}_f(\mathbf{x}_k) - \mathbf{y} \parallel$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \tau(\mathbf{x}_{k+1} - \mathbf{x}_k) \tag{15}$$

$$\lambda_k = \max(\lambda_{\mathrm{TR}}, \mathrm{R}_\lambda \lambda_\mathrm{k}) \tag{16}$$

  `end while`

- Update λ_{k+1}

$$\lambda_{k+1} = \max(\tau, \lambda_\mathrm{k}/\mathrm{r}_\lambda) \tag{17}$$

- Build response matrices

$$\begin{aligned}\Delta\mathbf{F_f}_{k+1} &= [\mathbf{f_f}(\mathbf{x}_{k+1}) - \mathbf{f_f}(\mathbf{x}_k), ..., \mathbf{f_f}(\mathbf{x}_{k+1}) - \mathbf{f_f}(\mathbf{x}_{max(k+1-n,0)})] \\ \Delta\mathbf{F_c}_{k+1} &= [\mathbf{f_c}(\mathbf{x}_{k+1}) - \mathbf{f_c}(\mathbf{x}_k), ..., \mathbf{f_c}(\mathbf{x}_{k+1}) - \mathbf{f_c}(\mathbf{x}_{max(k+1-n,0)})]\end{aligned}$$

- Update mapping

$$(\mathbf{U}_{\Delta\mathbf{F_f}}, \mathbf{U}_{\Delta\mathbf{F_c}}, \mathbf{\Sigma}_{\Delta\mathbf{F_f}}, \mathbf{\Sigma}_{\Delta\mathbf{F_c}}, \mathbf{V}) = \mathrm{GSVD}(\Delta\mathbf{F_f}_{k+1}, \Delta\mathbf{F_c}_{k+1}) \tag{18}$$

$$\mathbf{D} = \mathrm{diag}\left(\frac{\sigma_i^{\mathbf{F_c}} + \lambda_{k+1}(\sigma_1^{\mathbf{F_c}} + \tau)}{\sigma_i^{\mathbf{F_f}} + \lambda_{k+1}(\sigma_1^{\mathbf{F_f}} + \tau)}\right) \tag{19}$$

$$\mathbf{T_{k+1}} = \mathbf{U}_{\Delta\mathbf{F_c}}\mathbf{D}\mathbf{U}^T_{\Delta\mathbf{F_f}} \tag{20}$$

`end do`

In the manifold mapping scheme shown above, GSVD means generalized singular value decomposition[a] and σ_i are elements of the $\mathbf{\Sigma}$ matrices. λ, δ, τ,R_λ, r_λ, λ_{TR} and α are stabilization parameters controlling the amount of stabilization. In this work $\delta = 0$, $\tau = 0.1$, $R_\lambda = r_\lambda = 2$, $\lambda_{TR} = 1$ and $\alpha = 1 + 1 \cdot 10^{-10}$.

In case computationally expensive constraints, i.e., *fine* model constraints, are present, a mapping similar to the one given in Eq. (11) is created for each of the constraints. At each iteration a constrained optimization problem is solved with the fine models substituted by mapped coarse models. Details of the manifold mapping scheme with constraints can be found in Ref. 9.

Optimization problems

As stated earlier the temperature field is of importance to the microstructure evolution and residual stresses in the material being welded. In this work we consider two different formulations of a robust optimization problem, both with the aim of controlling the temperature field and reducing variations due to uncertainties in model parameters.

As a simplified method for controlling the temperature field the objective is to obtain a desired temperature in some point (ξ, η) on the plate. A succesful weld can roughly be described as having an average temperature, T_{Avg}, of $500^o\,C$ [29] under the tool. This temperature allows the material to become sufficiently soft for the the tool to stir it in order to create the weld. In the case of deterministic material parameters and design variables, i.e., without variations, this optimization problem could be formulated as given in Eq. (21)[16]

$$\begin{aligned}\min_{q,v} \quad & (T(\xi,\eta) - T_1)^2 \\ \text{s.t.} \quad & T_{Avg} = T_2\end{aligned} \tag{21}$$

where $T_1 = 573K$ and $T_2 = 773K$[b] are specified by the designer. However taking variations of material data or design variables into account the equality constraint cannot be satisfied for all values. Thus we change the formulation of the optimization problem to

$$\min_{q,v} || \{\mu_T \quad w_1\sigma_T \quad \mu_{Tavg} \quad w_2\sigma_{Tavg}\} - \{T_1 \quad 0 \quad T_2 \quad 0\} || \tag{22}$$

The formulation allows the inclusion of the standard deviations of the temperatures in the objective while at the same time aiming at specified mean values of the temperatures. The constants w_1 and w_2 allow the designer to emphasize minimization of standard deviations more or less compared to the mean responses. Furthermore it is directly applicable to the manifold mapping technique as given by Eq. (8). We recognize that this problem may be reformulated to fit into the form of Eq. (3).

However, the formulation in Eq. (22) does not allow the individual terms to be controlled. As mentioned earlier the average temperature under the tool is important for obtaining a succesful weld. Thus in order to

[a]Decomposition such that $\mathrm{GSVD}(\Delta\mathbf{F_f}, \Delta\mathbf{F_c})$ results in $\Delta\mathbf{F_f} = \mathbf{U}_{\Delta\mathbf{F_f}}\mathbf{\Sigma}_{\Delta\mathbf{F_f}}\mathbf{V}^T$ and $\Delta\mathbf{F_c} = \mathbf{U}_{\Delta\mathbf{F_c}}\mathbf{\Sigma}_{\Delta\mathbf{F_c}}\mathbf{V}^T$

[b]The temperatures used in this example are based on experience and may be adjusted to suit different welding scenarios, e.g. involving different materials.

have more control of the average temperature than the first formulation allows we reformulate the optimization problem to treat the mean value of the average temperature as well as the two standard deviations as constraints

$$\begin{aligned} \min_{q,v} \quad & ||\mu_T - T_1|| \\ \text{s.t.} \quad & \sigma_T \leq \sigma_{T_1} \\ & \mu_{Tavg} = \mu_{T_2} \\ & \sigma_{Tavg} \leq \sigma_{T_2} \end{aligned} \tag{23}$$

This formulation offers more control of the average temperature and the standard deviations.

Derivatives

As indicated in Eqs. (6) and (7) derivatives of the response functions with respect to uncertain parameters are required in order to evaluate the mean and standard deviations of the response. In this work we use both the first and second order Taylor expansions. For the fine (FE) model the derivatives of the responses can be obtained using a direct or an adjoint approach, see e.g. Refs. 5, 8.

We consider three different uncertain parameters; the conductivity k of the material, the heat input q and the welding speed v where the last two are design variables in the optimization problem. As discussed above derivatives of the temperature in a point as well as an average temperature is needed for the calculation of means and standard deviations. In case only first derivatives are considered this requires a total of 4 solutions of the equation system using the direct approach compared to 3 using the adjoint method.

Calculating second derivatives using the adjoint method requires knowledge of the field of first derivatives which can be obtained by the direct method leading to a mixed method using both direct and adjoint calculations. Calculating the second derivatives requires 10 solutions using the direct method compared to 6 for the mixed approach.

V. Examples: Uncertain material parameter

The optimization problem consired in this section is given by Eq. (22) considering only the conductivity k of the material as uncertain and $w_1 = w_2 = 1$. It is assumed that k is normally distributed with mean $\mu_k = 167\,\frac{W}{mK}$ and standard deviation $\sigma_k = 10\,\frac{W}{mK}$, i.e., a coefficient of variation of 6%. The design variables are the heat input q and the welding speed v and they are assumed deterministic, i.e., they take the values specified by the designer and do not vary from these values.

It should be emphasized that the results are obtained using a pure thermal model of the FSW process and are therefore not directly applicable to a real welding situation where also the mechanical effects of the process are important. However the results show that the method is useful for robust optimization.

The problem is solved with the manifold mapping technique using the model given by Eq. (1) implemented in Comsol Multiphysics as fine model and the analytical Rosenthal solution, Eq. (2), as coarse model. The coarse model optimization problems are solved using the built-in Matlab[19] simplex optimizer *fminsearch*. Only a first order Taylor expansion for calculation of mean and standard deviation is used in the coarse model. Using a second order Taylor expansion in the coarse model is also possible but to keep the coarse model as simple as possible we only use the first order expansion. Contrary, we show results using both first and second order Taylor expansions in the fine model. In Table 1 we show optimum values of heat input and welding speed for the case of deterministic optimization and robust optimization using the first and second order Taylor expansions in the fine model. In case of deterministic optimization the objective is reachable, i.e., it is possible to obtain $T(\xi, \eta) = T_1$ and $T_{Avg} = T_2$. We note that the fine model, i.e., the FE model, consists of a limited number of elements in order to reduce the computation time thereby allowing us to compare results to Monte Carlo simulations. A finer mesh model may change the results slightly.

Table 1. Optima for deterministic and robust optimization for the fine FE model and the coarse analytical model.

		q [W]	$v[\frac{mm}{s}]$
Fine FE model	Deterministic	1000.97	4.85
Fine FE model(a)	First order Taylor	1006.98	4.90
Fine FE model(b)	Second order Taylor	1003.86	4.88
Coarse analytical model	Deterministic	684.42	2.91
Coarse analytical model	First order Taylor	691.24	2.97

We see that the results for the fine model are very close in the sense that the differences in optimal welding speed and heat input for the three examples are small. The same is true for the coarse model. Although the optima for the fine and the coarse model are quite far apart they show a similar behaviour in the way that both welding speed and heat input increases from the deterministic optimum to the robust optimum. We should also note that even though the coarse and fine model optima are not close, the two models have the same behaviour meaning that an increase in heat input and a decrease in welding speed raises the temperatures and vice versa. Furthermore we note that the ratio between heat input and welding speed is almost the same for the two models, in the case of deterministic optimization approximately 206 $\frac{J}{m}$ for the FE model and 235 $\frac{J}{m}$ for the analytical model. Figure 2 show cross sectional temperature plots for the deterministic optima and for a robust optimum(see Section VI).

To compare robustness of the different optima given in Table 1, we do a robustness analysis of the two robust fine model design points based on the two Taylor expansions and Monte Carlo simulations. The results are given in Table 2.

Table 2. Robustness analysis of the optima given in table 1 using the first order and second order Taylor expansions and Monte Carlo simulation. (a) and (b) corresponds to the results in Table 1

Optimization method	First order Taylor (a)		Second order Taylor (b)	
	Taylor	Monte Carlo	Taylor	Monte Carlo
μ_T [K]	572.20	572.02	572.20	572.20
σ_T [K]	1.38	1.36	1.44	1.39
μ_{Tavg} [K]	773.16	773.51	773.16	773.16
σ_{Tavg} [K]	11.92	11.98	11.94	11.98

We note that the statistical data based on Taylor expansions are close to that obtained using Monte Carlo simulations. This suggests that for this problem, using Taylor expansions for the robust optimization will yield reliable results. The difference between using the first and second order Taylor expansion is small which indicates that for this particular problem one get good results using the first order expansion. However for larger uncertainties, i.e., larger standard deviations, it is expected that the second order expansion will yield more accurate results. Also note that the Monte Carlo simulations are performed using a limited number of samples (4000) so the results should not be considered exact values. However they show that the results from the Taylor expansions are sufficiently accurate for this type of engineering problem.

Figure 2 shows the optimized temperature field calculated using the fine model and second order Taylor expansion corresponding to the result (b) in Table 1.

VI. Examples: Uncertain material parameter and design variables

In the examples shown in the previous section only the conductivity k was treated as uncertain. Thus the design variables, the heat input and the welding speed, were assumed deterministic. However these may vary as well, due to uncertainties in e.g. the control of the machine. In this section we show examples of robustness optimization assuming that the mentioned variables, i.e., k, q and v, are uncertain.

Knowledge of the variations of different parameters may be hard to obtain in practice. For the example studied here we assume normally distributed and statistically independent variables with mean and standard deviations as shown in Table 3. We treat the standard deviation of the design variables as constants but one may define these as functions of the mean values. We only use the mean values of the heat input and welding speed as design variables as typically we will have no way of controlling the standard deviations of these.

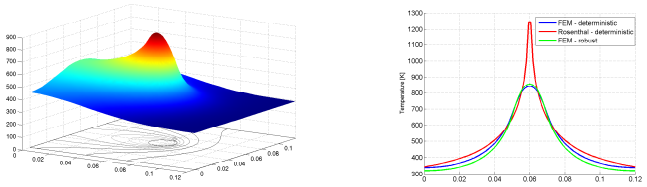

Figure 2. Left: Optimized temperature field for the example (b) in Table 1 using the second order Taylor expansion. z-axis value in K. Right: Temperature profile for cross section through the heat source, i.e., $\xi = 0$, for the fine(FEM) and the coarse(analytical) models in case of deterministic optimization. The results correspond to those given in Table 1. Also shown(green curve) is the temperature field for the robust optimization in case of uncertainties in k, q and v. This result corresponds to column two in Table 4.

However they could be included as design variables if for instance different machines would allow smaller values. In this case one should include a form of penalty term to the objective in order for the optimization not to require small standard deviations by default[14] .

Table 3. Mean and standard deviation of uncertain parameters. Mean values are not given for heat input and welding speed as they are treated as design variables.

	Mean	Standard deviation
Conductivity k $[\frac{W}{mK}]$	167	10
Heat input q $[W]$	-	50
Welding speed v $[\frac{mm}{s}]$	-	$1{\cdot}10^{-4}$

We again solve the optimization problem given by Eq. (22) and test the effect of using different weights w_1 and w_2. Second order Taylor expansions are used in the fine model. The results are shown in Table 4.

Table 4. Optimal design variables for the optimization problem Eq. (22) with one uncertain material parameter and two uncertain design variables. Also shown are statistical data based on second order Taylor expansions and percentage difference when compared to Monte Carlo simulations.

	$w_1 = w_2 = 1$	$w_1 = w_2 = 2$	$w_1 = w_2 = 3$	$w_1 = 1, w_2 = 5$
q^* [W]	1054.91	1177.30	1315.22	1528.46
v^* $[\frac{mm}{s}]$	5.25	6.16	7.19	8.80
μ_T [K]	566.92	554.09	540.04	519.16
σ_T [K]	13.93	11.93	10.30	8.72
μ_{Tavg} [K]	776.05	781.77	786.96	792.67
σ_{Tavg} [K]	26.24	23.93	21.74	18.97
Monte Carlo - Percent difference compared to Taylor expansions				
μ_T [K]	0	0	0	0
σ_T [K]	0.072	0.252	-0.097	0.344
μ_{Tavg} [K]	0	0	0	0.001
σ_{Tavg} [K]	-0.191	0.167	-0.506	0.633

The results shown in Table 4 show that increasing the weigths in the optimization problem has the desired effect meaning that the standard deviations of both the point temperature and the average temperature decrease. The cost of this is an increase in the difference between the mean temperatures and the desired values of these. We also note that the standard deviation of the average temperature is approximately twice as large as that for the point temperature. The decrease in standard deviations is obtained by increasing both heat input and welding speed and comparing to the result of the deterministic optimization given in Table 1 we see a large difference. As in the previous examples there is good agreement between statistical data based on Monte Carlo simulations and Taylor expansions. In the example with $w_1 = w_2 = 1$ the coefficient

of variation of the heat input is approximately 4.8% and 2% for the welding speed. Figure 2 shows the cross sectional temperature profile for the example from column two in Table 4 along with deterministic results from Table 1. We note the difference between the robust and the deterministic results for the fine FE model and see that the increased heat input and welding speed gives the robust result a higher maximum temperature and steeper temperature gradients next to the heat source.

Larger uncertainties

In the examples above the uncertainties were relatively small with coefficients of variation below 10%. Table 5 shows optimization results in case of larger variations.

Table 5. Optimal design variables for the optimization problem Eq. (22) with $w_1 = w_2 = 1$ and with one uncertain material parameter and two uncertain design variables. Also shown are statistical data based on Taylor expansions and percentage difference when compared to Monte Carlo simulations. The left column under each set of uncertainties corresponds to first order Taylor expansions while the right column corresponds to second order Taylor expansions.

	Uncertainties							
μ_k $[\frac{W}{mK}]$	167		167		167		167	
σ_k $[\frac{W}{mK}]$	10		20		20		30	
σ_q $[\frac{W}{m^3}]$	50		100		100		100	
σ_v $[\frac{mm}{s}]$	$1{\cdot}10^{-4}$		$2{\cdot}10^{-4}$		$5{\cdot}10^{-4}$		$5{\cdot}10^{-4}$	
	Optimization results							
q^* [W]	1057.84	1054.91	1180.04	1169.12	1218.57	1207.31	1232.08	1207.28
v^* $[\frac{mm}{s}]$	5.27	5.25	6.17	6.11	6.48	6.43	6.60	6.46
μ_T [K]	566.95	566.92	554.17	553.88	549.39	549.20	547.44	546.78
σ_T [K]	13.88	13.93	23.79	24.10	29.67	29.98	29.26	30.08
μ_{Tavg} [K]	776.04	776.05	781.74	781.86	782.37	782.42	782.24	782.41
σ_{Tavg} [K]	26.18	26.24	47.77	48.14	51.38	51.69	55.90	56.75
	Monte Carlo - Percent difference compared to Taylor expansions.							
μ_T	-0.025	-0.002	-0.164	-0.009	0.018	-0.011	-0.279	-0.059
σ_T	0.432	0.000	-1.471	0.290	1.348	-0.534	3.691	0.698
μ_{Tavg}	0.0490	0.001	0.157	0.005	0.284	0.000	0.437	-0.017
σ_{Tavg}	-0.500	0.343	-0.356	0.831	0.993	0.251	3.041	-0.247

It is seen that the first and second order approximations lead to optimum design points that are close to each other, but also that the difference between the optima for the two methods increases as the uncertainties become larger. Furthermore it can be seen that the difference between the statistical data based on first order Taylor expansions and that based on Monte Carlo simulations increases for larger uncertainties as would be expected. The second order Taylor expansions generally give results that are closer to that of Monte Carlo simulations, also for large uncertainties.

VII. Examples: Different formulation

In this section we show results obtained using the second formulation, Eq. (23), of the optimization problem. This will allow us to treat the mean of the average temperature and the standard deviations as constraints. As the average temperature is important to the weld quality this can be desirable. The optimization problem is solved using the manifold mapping technique for constrained problems[9] .

As an example of this method we assume the same uncertainties as shown in Table 3 and solve the problem for different values of the constraints. The results obtained using the second order Taylor expansion are shown in Table 6.

Table 6. Optimal design variables for the optimization problem given in Eq. (23) with one uncertain material parameter and two uncertain design variables.

Constraints				
σ_T [K]	20	20	20	10
μ_{TAvg} [K]	773	773	773	773
σ_{TAvg} [K]	25	22	20	10
Optimization results				
q^* [W]	1100.42	1253.75	1379.36	2682.59
v^* $[\frac{mm}{s}]$	5.69	6.98	8.05	18.88
μ_T [K]	557.40	536.20	520.53	416.08
σ_T [K]	12.87	10.54	9.27	6.85
μ_{Tavg} [K]	773.00	773.00	773.00	773.00
σ_{Tavg} [K]	25.00	22.00	20.00	10.00
Monte Carlo				
μ_T [K]	557.40	536.19	520.53	416.09
σ_T [K]	12.95	10.54	9.36	6.94
μ_{Tavg} [K]	773.01	773.00	772.99	773.00
σ_{Tavg} [K]	25.17	22.02	19.99	9.92

Clearly the objective value, given by $||\mu_T - T_1||$ and $T_1 = 573K$, increases as the constraints are tightened. We also see that the constraint on σ_T is not active in any of the four problems. This is in line with the previous results that showed that the standard deviation of T_{Avg} is larger than that of T. As before the optimization increases the heat input and welding speed in order to decrease the standard deviations while keeping the mean of the average temperature at the desired value. It is a general trend in the examples shown in this paper that in order increase robustness the heat input and the welding speed should be increased. Again good correpondance between Taylor expansions and Monte Carlo simulations is seen.

This example shows that the second formulation offers more control over the individual terms than the one considered in sections V and VI.

VIII. Computational efficiency

One of the motivations for using the manifold mapping technique for the robust optimization problem is that it does not require the user to supply gradients of the fine model objective and constraint functions. Furthermore it generally requires fewer fine model evaluations than methods that work directly on the fine model, see e.g. Ref. 9. To study this in more detail we compare the number of fine model evaluations required for different models for some of the examples shown in this paper. The fine model used in this paper consists of relatively few elements in order to allow us to validate results using Monte Carlo simulations. The number of fine model evaluations is therefore a better measure than actual CPU time when comparing the different methods as fine models used in practical problems will be much more computational expensive than the one used here. For expensive fine models the evaluation of these will typically be the dominant factor in the overall time consumption as the time spent on optimizing the coarse model is unaffected by the fine model size. The starting point in all cases is $q = 1000W$ and $v = 3\frac{mm}{s}$.

For the optimization problem given by Eq. (22) we compare the number of fine model evaluations using manifold mapping and the built-in Matlab[19] simplex optimizer *fminsearch* that solves unconstrained problems. The problem is similar to the one studied in Table 4 and the results are shown in Table 7.

Table 7. Comparison of number of fine model evaluations required using manifold mapping and simplex method for the examples shown in Table 4.

	Number of fine model evaluations	
	Manifold mapping	Simplex
Example 1	18	118
Example 2	22	122
Example 3	25	128
Example 4	19	152

Similarly in Table 8 we compare the number of fine model evaluations for the examples shown in Table 6 using the second formulation. As this is a constrained problem we compare to using SNOPT without supplying gradient information of the objective and constraint functions. Alternatively *fminsearch* could have been used by adding a penalty for infeasible design points.

Table 8. Comparison of number of fine model evaluations required using manifold mapping and SNOPT for the examples shown in Table 6.

	Number of fine model evaluations	
	Manifold mapping	SNOPT
Example 1	7	45
Example 2	8	44
Example 3	11	59
Example 4	8	54

It is seen that the manifold mapping technique requires significantly fewer fine model evaluations than the direct method in both cases. The results shown here were obtained by starting the optimization procedure in the same startpoint as used in manifold mapping. Starting in the coarse model optimum does not change the number of fine model evaluations significantly, and the conclusion that manifold mapping reduces the computational cost associated with evaluation of the fine model remains valid.

IX. Conclusion

In this work we studied the use of manifold mapping in connection to robust optimization of a thermal model of a FSW process. It was found that the use of Taylor expansions gave reliable statistical data when compared to results from Monte Carlo simulations for the tested uncertainties. Both first and second order expansions were used and, as expected, the second order expansions gave the best results for large uncertainties. The use of the manifold mapping technique was in part motivated by the fact that it does not require the user to provide gradient information of the objective and constraint functions. This is important in this case as we are using derivatives for the evaluation of the statistical data that are part of the objective and constraint functions. Gradient based optimization would require derivatives of the objective and constraint functions, i.e., derivatives of one order higher than what is used for the evaluation of statistical data. For the examples where the numbers of fine model evaluations were compared, manifold mapping proved to require significantly fewer fine models evaluations than direct optimization for both of the tested formulations.

Acknowledgments

This work is part of the InnoJoint project sponsored by the Danish Research Council for Technology and Production Sciences, contract number 274-05-0232.

References

[1]J.W. Bandler, R.M. Biernacki, S.H. Chen, P.A. Grobelny, and R.H. Hemmers. Space mapping technique for electromagnetic optimization. *IEEE Transactions on microwave theory and techniques*, 42, no. 12:2536–2544, 1994.

[2]J.W. Bandler, Q.S. Cheng, S.A. Dakroury, A.S. Mohamed, M.H. Bakr, K. Madsen, and J. Søndergaard. Space mapping: The state of the art. *IEEE Transaction on microwave theory and techniques*, 52, no. 1:337–361, 2004.

[3]H.G. Beyer and B. Sendhoff. Robust optimization - a comprehensive survey. *Computer methods in applied mechanics and engineering*, 197:3190–3218, 2007.

[4]C.M. Chen and R. Kovacevic. Finite element modeling of friction stir welding - thermal and thermomechanical analysis. *International Journal of Machine Tools and Manufacture*, 43:1319–1326, 2003.

[5]K.K. Choi and N.-H. Kim. *Structural Sensitivity Analysis and Optimization, Vol 1 & 2*. Springer-Verlag, New York, 2005.

[6]P.A. Colegrove, H.R. Shercliff, and P.L. Threadgill. Modelling and development of the trivex(tm) friction stir welding tool. *4th International Symposium on Friction Stir Welding*, 2003.

[7]Comsol. www.comsol.com, 2006.

[8]K. Dems. Sensitivity analysis in thermal problems - i: Variation of material parameters within a fixed domain. *Journal of thermal stresses*, 9:303–324, 1986.

[9]D. Echeverria. Two new variants of the manifold-mapping technique. *COMPEL The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, 26, No.2:334–344, 2007.

[10]D. Echeverria and P.W. Hemker. Space mapping and defect correction. *Computational methods in applied mathematics*, 5, no. 2:107–136, 2005.

[11]D. Echeverria and P.W. Hemker. Manifold mapping: a two-level optimization technique. *Accepted for publication in Computing and Visualization in Science*, 2008.

[12]F. Jurecka, K. Bletzinger, and J. Sauter. A metamodel update procedure for efficient robust design optimization. *Proceedings EuroMech Colloquium 482, London, Sep. 2007*, 2007.

[13]M.Z.H. Khandkar and J.A. Khan. Thermal modeling of overlap friction stir welding for al-alloys. *Journal of Materials Processing and Manufacturing Science*, 10:91–105, 2001.

[14]P.N. Koch, R.-J. Yang, and L. Gu. Design for six sigma through robust optimization. *Struct. Multidisc. Optim.*, 26:235–248, 2004.

[15]D. Lahaye, A. Canova, G. Grusso, and M. Repetto. Adaptive manifold-mapping using multiquadric interpolation applied to linear actuator design. *COMPEL The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, 26, no. 2:225–235, 2007.

[16]A.A. Larsen, M. Bendsøe, J. Hattel, and H. Schmidt. Optimization of friction stir welding using space and manifold mapping - an initial study of thermal aspects. *Structural and Multidisciplinary Optimization*, To appear.

[17]K. Lee and G. Park. A global robust optimization using kriging based approximation model. *JSME International Journal, series C*, 49, No. 3, 2006.

[18]S.N. Lophaven, H.B. Nielsen, and J. Søndergaard. DACE A Matlab Kriging Toolbox. *Technical Report IMM-TR-2002-12, IMM, Tech. University of Denmark*, 2002.

[19]Matlab. The mathworks, inc., 2006.

[20]M.D. McKay, R.J. Beckman, and W.J. Canover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21, No. 2:239–245, 1979.

[21]R.E. Melchers. *Structural Reliability Analysis and Prediction*. Ellis Horwood Ltd., 1987.

[22]M. Padulo, M.S. Campobasso, M.D. Guenov, and J. Maginot. Derivative-free uncertainty propagation methods for airfoil robust design. *Proceedings EuroMech Colloquium 482, London, Sep. 2007*, 2007.

[23]G-J. Park, T-H. Lee, K.H. Lee, and K-H. Hwang. Robust design: An overview. *AIAA Journal*, 44, No. 1:181–191, 2006.

[24]A. Parkinson. Robust mechanical design using engineering models. *Transactions of the ASME*, 117:48–54, 1995.

[25]D. Rosenthal. The theory of moving sources of heat and its application to metal treatments. *Transactions of the A.S.M.E.*, November, 1946.

[26]H. Schmidt and J. Hattel. CFD modelling of the shear layer around the tool probe in friction stir welding. *Friction Stir Welding and Processing III, San Francisco, USA, TMS*, 2005.

[27]H. Schmidt and J. Hattel. A local model for the thermomechanical conditions in friction stir welding. *Modelling and simulation in materials science and engineering*, Vol. 13:77–93, 2005.

[28]H. Schmidt, J. Hattel, and J. Wert. An analytical model for the heat generation in friction stir welding. *Modelling and Simulation in Materials Science and Engineering*, 12:143–157, 2004.

[29]C.C. Tutum, H. Schmidt, J. Hattel, and M.P. Bendsøe. Estimation of the welding speed and heat input in friction stir welding using thermal models and optimization. *Proc. 7th World Congress on Structural and Multidisciplinary Optimization (Ed. B.Y. Kwak et al), COEX Seoul, Korea, May 21-25, 2007 (CD-ROM, ISBN 978-89-959384-2-3 98550)*, pages pp. 2639–2646, 2007.

[30]TWI. The Welding Institute, www.twi.co.uk, 2007.

Appendix C

Article 3

Anders A. Larsen; Mathias Stolpe; Jesper Hattel; Henrik Schmidt. Optimization of Hardness in Friction Stir Welds. Proceedings of 8th World Congress on Structural and Multidisciplinary Optimization, 1-5 June 2009, Lisbon, Portugal. Edited by: H.C. Rodrigues, J.M. Guedes, F.R. Fernandes, J.O. Folgado and M.M. Neves. ISBN: 978-989-20-1554-5.

8th World Congress on Structural and Multidisciplinary Optimization
June 1-5, 2009, Lisbon, Portugal

Optimization of Hardness in Friction Stir Welds

Anders A. Larsen[1], Mathias Stolpe[1], Jesper Hattel[2] and Henrik Schmidt[2]

[1] Dept. of Mathematics, Tech. University of Denmark, Lyngby, Denmark, {a.a.larsen, m.stolpe}@mat.dtu.dk
[2] Dept. of Mech. Engineering, Tech. University of Denmark, Lyngby, Denmark, {jhat, hnbs}@mek.dtu.dk

1. Abstract
In this work we optimize a friction stir welding process with the aim of obtaining a desired hardness distribution in the welded plate. The hardness is a function of the temperature history in the plate which in turn depends on the process parameters. The optimization problem is solved using a manifold mapping technique in which the expensive 3D finite element model is used along with simpler, coarse models in order to reduce the number of expensive function evaluations. Two different coarse models are used; a simplified finite element model and a response surface.

2. Keywords: Friction stir welding, hardness, manifold mapping.

3. Introduction
Friction stir welding (FSW) is a solid state welding process that is well suited for aluminium welding. Unlike conventional welding methods the metal is not melted, which gives FSW a number of advantages, e.g. lower residual stresses and no need for inert gases making it an environmentally friendly process. The welding setup consists of the plates to be welded, the backingplate, and the tool, see figure 1. The welding is started by forcing the rotating tool into the material between the plates. When the tool probe is fully submerged and the tool shoulder is in contact with the plate surface it is moved forward along the weldline such that the material is heated due to friction and plastic deformations. The softened material from the two plates are stirred together by the mechanical action to form the weld.

Much work has been published on the thermal modelling of FSW with models ranging from simple analytical models based on Rosenthal solutions [11] to 3D numerical models. Traditionally the models are either transient Lagrangian, e.g. [16], or stationary Eulerian models, e.g. [13]. The first type allows finite plate dimensions to be used and allows the transient starting/stopping phases to be studied. The latter type has the advantages of fast solution times and the possibility to use a fine mesh close to the heat source. In the present work a 3D model consisting of the workpiece and the backing plate is used as the basis for calculation of the material hardness. An important aspect in modelling of friction stir welding is the heat source. In the real process the heat is generated due to friction and plastic deformation of the material close to the tool. In this work we use a temperature dependent heat source model from [13]. Using this model the heat input becomes part of the solution itself in contrast to more conventional heat sources where the total heat input is prescribed.

Some heat treated metals lose hardness and strength when subjected to thermal cycles such as welding. In [8, 9] a softening model for 6082-T6 aluminium is presented and applied to an example of a metal inert gas (MIG) weld using a Rosenthal solution for the thermal field. The results are compared to experimental values and good agreement is observed. A similar model is applied to friction stir welding of different aluminium alloys in [15]. In [12] the model is applied to FSW using a Rosenthal solution for calculating the temperature field. Also it is argued that the material in front of the tool should be fully softened in order to allow easy tool traversing, thereby limiting the welding speed. The hardness is proportional to the yield stress of the material and in [10] the material softening during welding is included in a residual stress calculation.

The present paper is organised as follows: Section 4 presents the thermal model and the hardness model, section 5 describes the optimization method used, section 6 shows the numerical results while section 7 concludes the paper.

4. Friction stir welding
This work focuses on optimizing the hardness distribution in friction stir welded plates. The model is divided into two parts; first a thermal analysis followed by calculation of the microstructure evolution

based on the temperature field.

4.1. Thermal model

The model is a 3D stationary Eulerian model consisting of the aluminium workpiece and the steel backing plate. The governing equation with a surface heat source, as defined in [13], is

$$\begin{aligned} 0 &= \nabla(-k\nabla T) + \rho c_p \mathbf{u}\nabla T \quad \text{in} \quad \Omega \\ q_{Heat\,source} &= \omega r(x,y)\tau(T) \qquad\quad \text{on} \quad \Gamma_{\text{Heat source}} \end{aligned} \tag{1}$$

where k is the thermal conductivity, ρ the material density, c_p the specific heat capacity, $\mathbf{u}$ the welding velocity vector, ω the rotational speed, $r(x,y)$ the distance from the tool axis of rotation to a given point, and $\tau(T)$ the temperature dependent shear yield stress. The tool is not modelled but the probe is omitted leaving a hole in the plate in which no heat transfer takes place. To include the rotational effect of the tool in the temperature field without solving for the actual mechanical deformations, a velocity field is enforced in a shear layer around the probe. The shear layer is a 1 mm thick layer around the probe. The model is sketched in Figure 1.

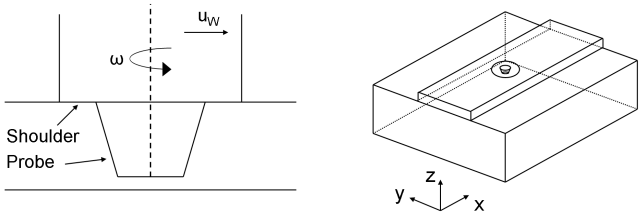

Figure 1: Left: Sketch of the tool during welding. Right: Sketch of the model showing backingplate, workpieces and probehole. Not to scale.

In the shear layer the velocity is given by the expression, [14].

$$\mathbf{u} = \left\{ \begin{array}{c} u \\ v \\ w \end{array} \right\} = \left\{ \begin{array}{c} -y\omega(1-\zeta) + \zeta u_w \\ x\omega(1-\zeta) \\ 0 \end{array} \right\} \tag{2}$$

where u_w is the translational speed and $\zeta \in [0,1]$ is a dimensionless parameter that gives the relative position inside the shear layer. Outside the shear layer, i.e. for $\zeta > 1$, the x-component of the velocity vector is $u = u_w$ while the other two components are zero. The parameter ζ is given by

$$\zeta = \frac{r(x,y) - R_{probe}}{w_{shear}} \tag{3}$$

where R_{probe} is the (z-dependent) radius of the probe and w_{shear} the thickness of the shear layer. On the probe surface $\zeta = 0$ meaning that the velocity equals the periferal velocity of the tool. By enforcing this no convective heat transfer is allowed through the probehole. The two-dimensional velocity field does not take any thickness-direction velocity into account. This would require a more detailed model, e.g. a computational fluid dynamics (CFD) model. Figure 2 shows an example of the velocity field for material entering the shear layer at different positions.

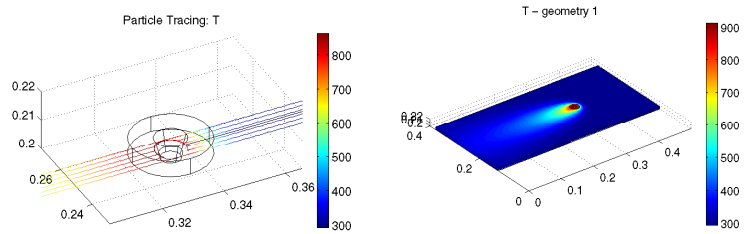

Figure 2: Left: The velocity field given by Eq. (2) with the color representing the temperature along the curve. The cone in the middle of the figure shows the probe hole while the outer cylinder shows the radius of the shoulder. Right: The temperature field in the plate.

The temperature dependent surface heat source is given by, [13],

$$q_{Heatsource} = \omega r(x,y)\tau(T) \tag{4}$$

This heat source is applied to the interface between the workpieces and the tool. The expression is obtained by integrating the (volume) heat generation due to plastic deformation during welding through the thickness of the shear layer. The model therefore takes the heat generation due to mechanical effects into account without solving for the full thermo-mechanical problem, hence the name thermal-pseudo-mechanical model. The temperature dependent shear yield stress $\tau(T)$ is a decreasing function of the temperature going to zero at the melting temperature. As $\tau(T)$ goes to zero so does the local heat generation thereby limiting the maximum temperatures that can be obtained.

During FSW the workpieces to be welded are placed on the steel backing plate. The heat transfer between the two parts is modelled with a convective heat transfer term, i.e. $q = h_{bottom}(T_{plate} - T_{backing})$. A value of $h_{bottom} = 700\frac{W}{m^2K}$, [13], is used.

4.2. Hardness evolution

A number of aluminium alloys experience softening when subjected to heat treatment such as welding. In this work we focus on welding of 6082-T6 plates and use the softening model described in [8] that allows calculation of the hardness after welding. The hardness is interesting as linear relations between the hardness and the yield and ultimate strength exist, [9].

The softening of the material during heat treatment is due to dissolution of hardening β'' (Mg_2Si) precipitates and the aim of the model is to determine the fraction of dissolved particles allowing the hardness to be calculated using linear interpolation between the original state and the fully dissolved state such that

$$\alpha = 1 - X_d = \frac{HV - HV_{min}}{HV_{max} - HV_{min}} \tag{5}$$

where α is the fraction of hardening precipitates, X_d the fraction of dissolved hardening precipitates, HV the hardness, HV_{max} the hardness of the material in T6 condition and HV_{min} the hardness of the fully softened material. For isothermal heat treatment at the reference temperature T_{r1} the material is fully softened, i.e. $X_d = 1$, after time t^*_{r1}. For a temperature T different from T_{r1} the time required to obtain full softening is given by

$$t^*_1 = t^*_{r1} \exp\left[\left(\frac{Q_s}{n_1 R} + \frac{Q_d}{R}\right)\left(\frac{1}{T} - \frac{1}{T_{r1}}\right)\right] \tag{6}$$

and the fraction of strengthening precipitates is

$$\alpha = 1 - \left(\frac{t}{t^*_1}\right)^{n_1} \tag{7}$$

Table 1 shows the values of the parameters used.

Table 1: Constants used in the hardness model, from [8].

Q_s	30 kJ/mol
Q_d	130 kJ/mol
n_1	0.5 (assumed constant)
t^*_{r1}	600 s
T_{r1}	375^o C
HV_{max}	110
HV_{min}	42
R	8.314 J/(molK)

For continuous heat treatments such as welding where the temperature is not constant the fraction of strengthening precipitates may be calculated by the integral

$$\alpha = 1 - \left(\int_0^{t_s} \frac{1}{t_1^*} dt \right)^{n_1} \tag{8}$$

where t_s is the duration of the heat treatment. Using eqs. (5), (6) and (8) the hardness may be calculated for a given thermal cycle. Note that t_1^* is a function of the temperature in the plate and the integration in eq. (8) is not trivial. For the FSW model described in the previous section this means that the integral should be calculated along a streamline following the imposed velocity field. As the material close to the tool is heated extensively while the material further away only experience a slight increase in temperature the characteristic hardness profile seen in Figure 4 is obtained.

For long, high temperature thermal cycles the model predicts hardness below HV_{min} as the α values decreases below zero. In this case α is set to zero thus limiting the minimum hardness to HV_{min}. However, when doing optimization it seems advantageous not to impose this limit on α as this results in large areas in variable space (u_w, ω) where $HV = HV_{min}$.

It should be noted that the above hardness model gives the hardness immediately after welding. Subsequent natural ageing will restore some of the hardness loss in the softened zone but this is not taken into account in this model.

5. Optimization

The purpose of the optimization problems considered in this work is to control the hardness distribution in the welded plate. This means controlling the width of the low-hardness zone and the transition from low hardness to high hardness.

The main process parameters in FSW are the welding speed u_w and the rotational speed ω and they are used as optimization variables in this work. Also heat transfer coefficients to the air and, mainly, between the workpieces and the backingplate, i.e. h_{bottom}, influence the temperature fields but these parameters are not easily controlled and therefore not included as process variables.

5.1. Manifold mapping

In this work a manifold mapping technique, see e.g. [1, 7], is used with the intention of reducing the computation time. In manifold mapping a mapping $\mathbf{S}$ is established between the response of two different models; the *fine* model to be optimized and the *coarse* model that is less accurate, but faster to evaluate than the fine model. The idea is to use the mapped coarse model as a surrogate for the fine model. As the mapping is gradually improved the mapped coarse model will resemble the fine model ever closer.

The optimization problem is formulated as a minimization of the norm of the difference between the fine model response and the desired values of the response function $\mathbf{y}$,

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{f_f}(\mathbf{x}) - \mathbf{y}||. \tag{9}$$

It should be noted that $\mathbf{f_f}$ and $\mathbf{y}$ are vectors of dimension m and in many cases $\mathbf{y} = \mathbf{0}$. The mapping $\mathbf{S}$ that maps $\mathbf{f_c}(\mathbf{x}^*)$ to $\mathbf{f_f}(\mathbf{x}^*)$ is defined as

$$\mathbf{S}(\mathbf{f_c}(\mathbf{x})) = \mathbf{f_f}(\mathbf{x}^*) + \bar{\mathbf{S}}(\mathbf{f_c}(\mathbf{x}) - \mathbf{f_c}(\mathbf{x}^*)) \tag{10}$$

where $\bar{\mathbf{S}}$ is determined by the coarse and fine model Jacobians such that $\bar{\mathbf{S}}$ maps the coarse model Jacobian to the fine model Jacobian, i.e. $\bar{\mathbf{S}}\mathbf{J_c}(\mathbf{x}^*) = \mathbf{J_f}(\mathbf{x}^*)$ or $\bar{\mathbf{S}} = \mathbf{J_f}(\mathbf{x}^*)\mathbf{J_c}^\dagger(\mathbf{x}^*)$ where † denotes the pseudoinverse.

By approximating the Jacobians with difference matrices at each iteration k, such that

$$\bar{\mathbf{S}}_k = \Delta\mathbf{F}_{\mathbf{f}k}\Delta\mathbf{F}_{\mathbf{c}k}^{\dagger} \tag{11}$$

the fine model Jacobian is not used and only fine model function values are required. $\Delta\mathbf{F_f}$ and $\Delta\mathbf{F_c}$ are matrices containing changes in fine and coarse model responses obtained in previous iterations

$$\begin{aligned}\Delta\mathbf{F}_{\mathbf{f}k+1} &= [\mathbf{f_f}(\mathbf{x}_{k+1}) - \mathbf{f_f}(\mathbf{x}_k), ..., \mathbf{f_f}(\mathbf{x}_{k+1}) - \mathbf{f_f}(\mathbf{x}_{max(k+1-n,0)})] \\ \Delta\mathbf{F}_{\mathbf{c}k+1} &= [\mathbf{f_c}(\mathbf{x}_{k+1}) - \mathbf{f_c}(\mathbf{x}_k), ..., \mathbf{f_c}(\mathbf{x}_{k+1}) - \mathbf{f_c}(\mathbf{x}_{max(k+1-n,0)})]\end{aligned}$$

Using the mapping, problem (9) is substituted by the sequence of surrogate problems

$$\begin{aligned}\mathbf{x}_{k+1} &= \underset{\mathbf{x}}{\text{argmin}}||\mathbf{S}_k(\mathbf{f_c}(\mathbf{x})) - \mathbf{y}|| \\ &= \underset{\mathbf{x}}{\text{argmin}}||\Delta\mathbf{F}_{\mathbf{f}k}\Delta\mathbf{F}_{\mathbf{c}k}^{\dagger}(\mathbf{f_c}(\mathbf{x}) - \mathbf{f_c}(\mathbf{x}_k)) + \mathbf{f_f}(\mathbf{x}_k) - \mathbf{y}||\end{aligned} \tag{12}$$

where the optimization is performed using the mapped coarse model that is fast to evaluate and where, in this case, gradients can be calculated efficiently. In short the manifold mapping scheme consists of the following steps: optimize mapped coarse model; evaluate fine model in the new point; update the mapping and repeat.

For optimization with constraints a similar mapping is done for these, see [1] where also a version, for unconstrained problems, including stabilization is presented. The stabilization is introduced to limit step lengths for ill-conditioned problems and takes effect when a step increases the objective function value. In this case the stabilization factor is increased until a succesful step, i.e one that reduces the objective function value, is taken. Experience with the (constrained) optimization problems at hand suggested that a method of controlling step sizes should be included. This is obtained through the use of a filter, [3, 4], and a trust region approach.

The idea behind a filter is that the aim of the optimization is minimizing the objective function while satisfying the constraints, or minimizing the objective function $F_f(\mathbf{x})$ and minimizing constraint violation $h(\mathbf{C}_f(\mathbf{x}))$, where $\mathbf{C}_f(\mathbf{x})$ is the fine model constraint function. The filter is used to decide whether or not a step is acceptable and how to update the trust region radius. In short, the filter is a list of points, objective function value and constraint violation, such that no point i dominates another point j in the sense that

$$F_f^i < F_f^j \quad \text{and} \quad h^i < h^j \tag{13}$$

Every new (*fine* model) point, $F_f(\mathbf{x}_k)$ and $h(\mathbf{C}_f(\mathbf{x}_k))$, that is genererated during the optimization is tested for acceptability to the filter and if acceptable it is included. Whenever a new point is included the filter is updated such that any dominated points are removed. The constraint violation is $h(\mathbf{C}_f(\mathbf{x})) = \sum_{j=1}^{m}\max(0, c_j)$ where $c_j > 0$ if constraint j is violated.

In the present work where we are using the manifold mapping technique the filter consists of points related to the *fine* model. In case a point is *not* acceptated by the filter we shrink the trust region and repeat the optimization of the mapped coarse model. If a point is accepted we calculate the ratio of the change in the actual fine model objective function value to the change in the mapped coarse model objective function value. If this ratio is close to 1 we increase the trust region size. The use of a filter-technique in combination with surrogate based optimization is briefly discussed in [2].

5.2. Coarse models

In manifold mapping a coarse model is used along with the fine model described in a previous section. Different coarse models may be used; in this work we use a greatly simplified version of the fine model as well as a radial basis function response surface.

The first coarse model is a 3D finite element model similar to the fine model but simplified by using a very coarse mesh and omitting the rotational velocity field and the backing plate. By not including the velocity field the hardness calculation becomes significantly faster as the integration is performed along straight lines. Not modelling the backingplate removes many degrees of freedoms from the system thereby making the solution procedure much faster. Analytical expressions for gradients of objective and constraint functions with respect to the process variables can be obtained for the (mapped) coarse model such that a gradient based optimization method may be used. In this work SNOPT, [5], is used for solving all coarse model optimization problems.

The response surface method used here is based on an augmented radial basis function technique, [6], in which the response f_c in some point $\mathbf{x}$ is given by

$$\begin{aligned} f_c(\mathbf{x}) &= \sum_{i=1}^{N} \lambda_i \phi(r) + \sum_{j=1}^{M} P_j(\mathbf{x}) b_j \\ r &= \|\mathbf{x} - \mathbf{x}_i\|_2 \\ P_j &= \begin{bmatrix} 1 & x & y & x^2 & xy & ... \end{bmatrix} \end{aligned} \tag{14}$$

The constants λ and b are obtained by solving the equation system

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \boldsymbol{\lambda} \\ \mathbf{b} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F} \\ \mathbf{0} \end{Bmatrix} \tag{15}$$

where the $N \times N$ matrix A has elements $A(i,j) = \phi\left(\|\mathbf{x}_i - \mathbf{x}_j\|_2\right)$ and the $N \times M$ matrix has elements $B(i,j) = P_j(\mathbf{x}_i)$. The vector $\mathbf{F}$ contains the function values corresponding to the datapoints $\mathbf{x}_i$ while vectors $\boldsymbol{\lambda}$ and $\mathbf{b}$ contains the unknown coefficients λ and b.

Several different radial basis functions ϕ may be used, however in this work we use $\phi(r) = \left(r^2 + h^2\right)^{\frac{1}{2}}$ where h is a parameter that controls the smoothness of the surface and can be adjusted by the designer. For the polynomial term in eq. (14) we choose $M = 1$ such that only a constant term is added.

The advantage of the first coarse model, i.e. the finite element model, compared to the response surface model is that it is a physically based model meaning that it is expected to have similar behaviour as the fine model all over the parameter space. However, each evaluation of this model is quite time consuming such that the coarse model optimization problem becomes expensive to solve compared to when using the response surface method. The response surface on the other hand is expensive to create in the first place but the subsequent coarse model optimization problems can be solved extremely fast. For problems where the approximate location of the optimum is not known the response surface must span a large variable space which may result in reduced accuracy or that more points should be used. When using manifold mapping with a response surface as coarse model it is possible to add the fine model evaluation performed in each iteration to the response surface and in that way iteratively update both the mapping and the coarse model. That option has not been used in this work.

6. Examples

The aim of the optimization problems considered here is to control the size and shape of the softened zone, including the transition from the hard to the soft zone. Based on tensile test simulations of the welded material it seems to be desirable not to have large gradients in the hardness and yield strength. Also we require that the material in front of the tool should be sufficiently softened and that the width of the low-hardness zone is not too large. Mathematically we aim at solving the problem

$$\begin{aligned} \min_{u_w, \omega} \quad & \|\Delta\mathbf{HV} - \mathbf{y}\| \\ \text{s.t.} \quad & \sigma_y(\mathbf{x}_A) \leq 10\, MPa \\ & HV(\mathbf{x}_B) \geq 0.85 \cdot HV_{max} \\ & P_{input} \leq 5\, kW \end{aligned} \tag{16}$$

where $\Delta\mathbf{HV} = [HV(\mathbf{x}_2) - HV(\mathbf{x}_1) \quad HV(\mathbf{x}_3) - HV(\mathbf{x}_2) \quad HV(\mathbf{x}_4) - HV(\mathbf{x}_3) \quad HV(\mathbf{x}_5) - HV(\mathbf{x}_4)]$ contains differences in hardness measured in four intervals perpendicular to the welding direction, i.e. in the y-direction, and $\mathbf{y}$ is a zero vector. The objective is thus to minimize the difference in hardness between the five measurement points $\mathbf{x}_i$, $i = 1, ..., 5$. An abrupt change in hardness between just two of the points yields a high objective function value and therefore conflicts with the goal of reducing the gradient in the hardness. We notice that the objective function takes its minimum value 0 if the hardness is equal in all points, e.g. $HV(\mathbf{x}_i) = HV_{min}$ or $HV(\mathbf{x}_i) = HV_{max}$ for $i = 1, ..., 5$. The first constraint in eq. (16) requires that the material yield stress in a point $\mathbf{x}_A$ in front of the tool is reduced such that the tool is easily traversed and reducing the risk of probe breakage. The yield stress depends on both the instantenous temperature as well as the hardness based on the thermal history $\sigma_y = (1 - \frac{T - T_0}{T_m - T_0})(\sigma_{y,min} + \alpha(\sigma_{y,max} - \sigma_{y,min}))$ where α is the strength parameter in Eq. (8) and

$\sigma_y(MPa) = 3.0HV - 48.1$ is related to the hardness, [9]. T_0 and T_m is the reference temperature for the yield stress and the solidus temperature, respectively. The second constraint controls the maximum width of the soft region by requiring that the hardness in a point $\mathbf{x}_B$ outside the tool region is above 85% of the maximum hardness. This constraint removes the possibility of $HV(\mathbf{x}_i) = HV_{min}$, $i = 1, .., 5$. The last constraint limits the maximum heat input to the workpiece as limited by the welding machine. The total heat input is found by integration of expression (4) over the tool surface.

6.1. Results

Below we present results for the optimization using the two different coarse models. First, results obtained using the simplified FE model as coarse model is shown in table 2.

Table 2: Results of optimization. The coarse model was a simplified finite element model.

Obj. value (Coarse model optimum)	Obj. value (final)	u_w	ω	Fine model evals.
61.6	44.9	8.90 $\frac{mm}{s}$	3228 RPM	10

In the optimum the first and third constraints are active. The yield stress in front of the tool is reduced by a combination of high temperature in the point and reduction due to microstructural changes. In this example $(1 - \frac{T-T_0}{T_m-T_0}) = 0.1$ and $\alpha = 0.11$ which combine to give a yield stress of 10 MPa. It seems as if the optimizer is trying to increase the welding speed in order to reduce the width of the softened zone and increase the overall hardness level in order to improve the objective function value. In order to satisfy the constraint on σ_y in front of the tool the rotational speed must increase as well to soften the material sufficiently. The power constraint therefore becomes the limiting factor. Increasing the allowable power will reduce the objective function value by increasing welding speed and rotational speed.

Figure 3 shows the objective function value and the constraint value for the two active constraints as function of iteration number.

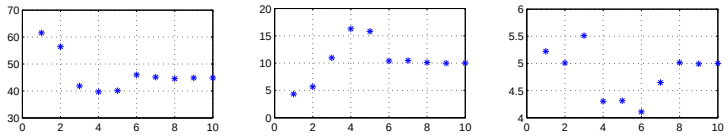

Figure 3: Objective and constraint function values as function of iteration number. Left: Objective function value. Center: Constraint on σ_y in front of tool. Right: Constraint on power.

We observe that the objective function value clearly decreasess during the iterations. For iteration 1 the fine model is evaluated in the coarse model optimum and it is seen that the objective value in this point is significantly higher than in the fine model optimum at iteration 10. Also it is interesting to note that the coarse model optimum is not feasible, see the right plot in Figure 3, so that in the first iteration the optimizer minimizes the constraint violation of the coarse model. As the mapping is improved the optimum of the mapped coarse model changes and the problem becomes feasible. The coarse model optimum is found at a much higher rotational speed than the fine model optimum which may be explained by the fact that the backing plate is not modelled as in the fine model. The heat loss from the bottom of the workpiece is given by $q = h_{bottom}(T - T_2)$ where T_2 in the fine model is the backing plate temperature while in the coarse model T_2 is set to the ambient temperature that is much lower than the true backing plate temperature. Thus, the coarse model requires a higher value of ω in an attempt to compensate for the larger heat loss.

From Figure 3 we see that the two constraints are working against each other in the sense that a high power value corresponds to a warm weld which in turn reduces the yield stress in front of the tool. Similarly a low power value and high yield stress are connected.

Table 3 shows the corresponding results when using the response surface as coarse model.

Table 3: Results of optimization. The coarse model was a response surface model. * The number in the paranthesis is the number of fine model evaluations used to create the response surface.

Obj. value (Coarse model optimum)	Obj. value (final)	u_w	ω	Fine model evals.*
41.2	44.9	8.90 $\frac{mm}{s}$	3228 RPM	19(9)

First, we notice that the optimal variables are identical to the previous example using the finite element coarse model. In the present case the objective function value increases during the iterations as the fine model evaluated in the coarse model optimum is not feasible. However, the response surface coarse model is a better approximation to the fine model than in the example with the coarse finite element model in the sense that the coarse model optimum is closer to the fine model optimum. The response surface used here was created using a regular grid of 3×3 data points but more points may of course be used to create a better approximation and that way maybe reduce the number of manifold mapping iterations required.
In Figure 4 the hardness profile in the cross section of the plate.

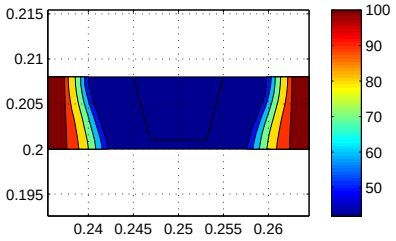

Figure 4: Hardness profile in the plate after welding. The probe is sketched in the center.

We note that the hardness profile is almost symmetric with only a slight non-symmetry due to the rotational velocity field. Furthermore the low hardness zone is widest in the top due to the heat generation from the tool shoulder. Points outside the radius of the shoulder have hardnesses very close to the base material and the hardness contours roughly follow the temperature contours as seen in a cross section at the tool.

In a real weld the risk of void creation in the weld line could be a limiting factor on the welding speed. However, to take this effect into account would require a model that calculates the material flow in contrast to the prescribed flow field used here.

7. Conclusions

In this paper a model of a friction stir welding process was optimized with the objective of contolling the material hardness. This was done using a manifold mapping technique and two different coarse models. The optimal process variables for the two methods were identical while the response surface coarse model required more fine model evaluations due to the need for creating the response surface. It was found that the optimal welding speed and, especially, the rotational speed were large compared to values traditionally used. To reduce the objective function it is desirable to weld fast, while the rotational speed is increased in order to generate enough heat to satisfy the constraints on the material yield stress in front of the tool. The combined effect is that the total power increases such that the the power constraint in the optimization problem becomes active.

8. Acknowledgements

This work is part of the InnoJoint project sponsored by the Danish Research Council for Technology and Production Sciences, contract number 274-05-0232.

9. References

[1] Echeverria, D.: Two new variants of the manifold-mapping technique. COMPEL The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, 2007, vol. 26, No. 2, pages:334-344

[2] Eldred, M.S. and Dunlavy, D.M.: Formulations for surrogate-based optimization with data fit, multifidelity, and reduced-order models. 11th AIAA/ISSMO Multidisciplinary analysis and optimization conference, 6-8 September 2006, Portsmouth, Virginia.

[3] Fletcher, R. and Leyffer, S.: Nonlinear programming without a penalty function. Math. Program, Ser. A, 2002, vol. 91, pages: 239-269

[4] Fletcher, R.; Leyffer, S. and Toint, P.L.: On the global convergence of a filter-sqp algorithm. SIAM J. Optim, 2002, vol. 13, no. 1, pages: 44-59

[5] Gill, P.E.; Murray, W. and Saunders, M.A.: Users Guide for SNOPT Version 7: Software for Large-Scale Nonlinear Programming. 2006

[6] Krishnamurthy, T.: Response surface approximation with augmented and compactly supported radial basis functions. 44th AIAA/ASME/ASCE/AHS Structures, Structural Dynamics, and Materials Conference, 7-10 April 2003, Norfolk, Virginia

[7] Larsen, A.A.; Lahaye, D.; Schmidt, H.; Hattel, J. and Bendsøe, M.: Robust Optimization of Thermal Aspects of Friction Stir Welding Using Manifold Mapping Techniques. 12th AIAA/ISSMO Multidisciplinary analysis and optimization conference, Victoria, British Columbia, sep. 2008

[8] Myhr, O.R. and Grong, Ø, Process modelling applied to 6082-T6 aluminium weldments-I. Reaction kinetics. Acta metall. mater., 1991, vol. 39, no. 11, pp. 2693-2702

[9] Myhr, O.R. and Grong, Ø, Process modelling applied to 6082-T6 aluminium weldments-II. Applications of model. Acta metall. mater., 1991, vol. 39, no. 11, pp. 2703-2708

[10] Richards, D.G.; Prangnell, P.B.; Williams, S.W. and Withers, P.J.: Global mechanical tensioning for the management of residual stresses in welds. Materials science & engineering A, 2007, vol. 489, pages: 351-362

[11] Rosenthal, D.: The Theory of Moving Sources of Heat and Its Application to Metal Treatments. Transactions of the A.S.M.E., 1946, November.

[12] Russell, M.J. and Shercliff, H.R.: Analytical modelling of microstructure development in friction stir welding. Proc. 1st Int. Symp. On Friction Stir Welding, Thousand Oaks, California, USA, June 1999.

[13] Schmidt, H.B. and Hattel, J.H.: modelling in friction stir welding. Scripta Materialia 58 (5) 332-337,2008

[14] Schmidt, H.B. and Hattel, J.H.: Modelling heat flow around tool probe in friction stir welding. Science and Technology of Welding and Joining, 2005, vol. 10, no. 4

[15] Shercliff, H.R.; Russell, M.J.; Taylor, A. and Dickerson, T.L.: Microstructural modelling in friction stir welding of 2000 series aluminium alloys. Mecanique & industries, 2005, vol. 6, pages: 25-35

[16] Soundararajan, V.; Zekovic, S. and Kovacevic, R.: Thermo-mechanical model with adaptive boundary conditions for friction stir welding of Al 6061. International journal of machine tools & manufacture, 2005, vol. 45, pages:1577-1587

Appendix D

Article 4

To be submitted.

Estimating the workpiece-backingplate heat transfer coefficient in friction stir welding using an inverse modelling technique

Anders A. Larsen, Mathias Stolpe, Henrik Schmidt and Jesper Hattel

Abstract

In this work inverse modelling combined with continuous optimization techniques is used to determine the heat transfer coefficient between the workpiece and the backingplate in a friction stir welding process. The heat transfer from workpiece to backingplate is important for the temperature field in the workpiece and therefore also the mechanical properties of the welded plate. The contact pressure between the two parts is highly non-uniform leading to a non-uniform heat transfer coefficient. Experimental measurements of temperatures in different positions in the workpiece are used together with an Eulerian 3D stationary thermal model of the process in order to determine the position dependent heat transfer coefficient.

1 Introduction

Friction stir welding (FSW) is a solid state welding process in which a rotating tool is used for joining of the two workpieces. The tool consist of a cylindrical probe that is forced into the weldline between the plates to be welded. The probe is attached to a cylinder of larger diameter that is in contact with the top surfaces of the plates, see Figure 1. The rotation of the tool generates heat by friction and plastic deformation of the plate material, thereby softening the workpiece material close to the tool. The maximum temperature is below the solidus temperature such that no molten phase is present. During the translational movement of the tool, the plate material is forced to flow in a complex pattern around the probe and that way material from the two plates merges to form the weld. During welding, the plates are clamped to a backingplate to restrain their movement and it is the thermal effects of the workpiece-backingplate contact that is the focus of this work.

A large number of different models of the FSW process may be used for modelling different aspects of the process, e.g. thermal models [2] or material flow models [3, 11]. In this work we study the heat transfer at the workpiece-backingplate contact interface and therefore restrict ourselves to a 3D Eulerian steady state thermal model.

The workpiece-backingplate contact is responsible for the majority of the heat loss from the workpiece during FSW and is discussed in some detail in [13]. Typical values of heat transfer coefficients used in the literature are $h_{air} \approx 10\,\frac{W}{m^2K}$ for the air-workpiece heat transfer, and $h_{bottom} \approx 1000\,\frac{W}{m^2K}$ for the workpiece-backingplate heat transfer [12, 8] with values ranging from $h_{bottom} = 350\,\frac{W}{m^2K}$ in [2] to $h_{bottom} = 5000\,\frac{W}{m^2K}$ in [9]. In most published work h_{bottom} is assumed uniform although it is expected to be largest in areas of high contact pressure, i.e. under the tool. In [13] a pressure dependent h_{bottom} is determined and it is found to be highly non-uniform with a very high value under the tool and very small values in areas in front of the tool. In [10] inverse modelling is used to determine a uniform value of h_{bottom} from experiments. The problem is formulated as a continuous optimization problem, and a genetic algorithm is used for minimizing an objective function based on the differences between measured and calculated temperatures. In [8] different values of h_{bottom} are tested and the predicted temperatures compared to measurements. A non-uniform distribution of h_{bottom} taking values of $100000\frac{W}{m^2K}$, $10000\frac{W}{m^2K}$ and $1000\frac{W}{m^2K}$ in regions under the tool, behind the tool, and in other regions, respectively. In this work we aim at using gradient-based optimization techniques to determine h_{bottom} from experimental temperature measurements. As stated above, the heat transfer coefficient between the workpiece, or backingplate, and the air is in general much smaller than h_{bottom} and we therefore restrict our attention to the latter. There are, however, no principal difficulties involved in including h_{air} in the inverse analysis.

2 Thermal model

The 3D thermal model of the process consists of two parts; the aluminium workpiece and the steel backingplate, and the model is a steady state Eulerian model, see Figure 1 for a sketch. The tool is not modelled but instead a volume corresponding to the tool probe is cut away from the workpiece. The heat equation to be solved in both the workpiece and the backingplate is

$$0 = \nabla(-k\nabla T) + \rho c\mathbf{u}\nabla T \tag{1}$$

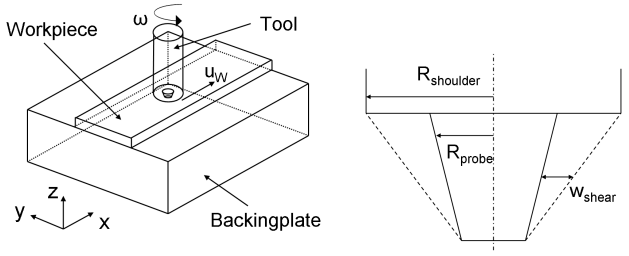

Figure 1: Left: Sketch of the welding setup showing the tool, the workpiece and the backingplate. Note that the tool is not included in the thermal model. Right: Sketch of the tool with the shear layer where the rotational velocity field is enforced. Not to scale.

where k is the thermal conductivity, $[\frac{W}{mK}]$, ρ the material density, $[\frac{kg}{m^3}]$, c the specific heat capacity,$[\frac{J}{m^3K}]$, and $\mathbf{u}$ the welding velocity vector, $[\frac{m}{s}]$. In order to include the rotational effect of the tool and material flow on the temperature field, a (planar) rotational velocity field is enforced in a shear layer around the probe. The velocity vector $\underline{u}$ is, for $0 \leq \zeta < 1$, modelled as

$$\underline{u} = \left\{ \begin{array}{c} -y\omega(1-\zeta) + \zeta u_w \\ x\omega(1-\zeta) \\ 0 \end{array} \right\} \tag{2}$$

where $\zeta = \frac{r - R_{probe}}{w_{shear}}$ is the dimensionless position inside the shear layer defined as a cone stretching from the probe tip to the outer edge of the shoulder, see Figure 1. w_{shear} is the z-dependent thickness of the shear layer. u_w is the translational welding speed, $[\frac{m}{s}]$, and ω is the rotational speed, $[\frac{rad}{s}]$. For points outside the shear layer the velocity vector only has the linear welding speed component, i.e.

$$\underline{u} = \left\{ \begin{array}{c} u_w \\ 0 \\ 0 \end{array} \right\} \tag{3}$$

The surface heat source is modelled using the thermal-pseudo-mechanical (TPM) model, [12], that uses a temperature dependent shear yield strength $\tau(T)$. The local heat generation is given by

$$q(T, x, y) = \omega r(x, y)\tau(T) \tag{4}$$

where $r(x, y)$ is the radial distance from the axis of rotation to a given point (x, y). This heat source is applied to the contact surfaces, denoted $\Gamma_{heat\,input}$, between the tool and the workpiece. The tool itself is not modelled but to take heat loss into the tool into account the term $h_{tool}(T - T_\infty)$ is added to the contact surfaces between the workpiece and the tool. h_{tool} contains the heat transfer coefficient from the tool to the machine and the 1D heat flow through the tool. For the specific tool $h_{tool} = 640\ \frac{W}{m^2K}$ which, in the examples below, corresponds to that approximately 10% of the total heat generation goes into the tool. This figure seems to be in agreement with figures from the literature, [5], although this is a simplified analysis of the heat loss to the tool.

The focus of this work is the heat transfer from the workpiece to the backingplate through the contact surface denoted Γ_{bottom}. This heat transfer is modelled with the expression

$$q_{bottom}(x, y, \underline{\xi}) = h_{bottom}(x, y, \underline{\xi})(T(x, y, \underline{\xi}) - T_2(x, y, \underline{\xi})) \tag{5}$$

where we emphasize the spatial dependence and the dependence on $\underline{\xi}$. $\underline{\xi}$ is a vector of the optimization variables whose values are to be optimized in the optimization problems. In the examples presented later, $\underline{\xi}$ contains between 1 and 6 elements that control the spatial distribution of h_{bottom}. T is the workpiece temperature and T_2 is the backingplate temperature. It is the (x, y) variation of h_{bottom} that we wish to determine, but to ease notation we write h_{bottom} for $h_{bottom}(x, y, \underline{\xi})$ in the following.

Solving the coupled governing equations for the workpiece and the backingplate gives T and T_2, i.e. the workpiece and backingplate temperature fields. These may be compared to pointwise experimental values obtained from [12]. Based on the governing equations it is possible to set up equations, using either a direct or an adjoint method, for derivatives of the temperature field with respect to the optimization variables, i.e $\frac{\partial T}{\partial \xi_i}$, for use in gradient based optimization methods. The thermal model is solved using Comsol Multiphysics [4] and the material and process parameters used are shown in table 1.

Table 1: Material and process parameters used in the thermal model.

u_w	$0.66\ \frac{mm}{s}$	
ω	$56.03\ \frac{rad}{s}$	
h_{air}	$10\ \frac{W}{m^2K}$	
	Workpiece	Backingplate
k	$130\ \frac{W}{mK}$	$44.5\ \frac{W}{mK}$
ρ	$2810\ \frac{W}{mK}$	$7850\ \frac{W}{mK}$
c	$960\ \frac{J}{kgK}$	$475\ \frac{J}{kgK}$

Sensitivity analysis

The governing equations for the thermal model, as presented above, may be written in weak form as

$$0 = \int_\Omega \nabla w k \nabla T + \int_\Omega w \rho c_p \underline{u} \nabla T + \int_{\Gamma_{bottom}} w h_{bottom}(T - T_2) + \int_{\Gamma_{air}} w h_{air}(T - T_\infty) - \int_{\Gamma_{heat\ input}} w\left(\omega r \tau(T) - h_{tool}(T - T_\infty)\right)$$

where w is a test function, and Ω is the computational domain. Γ_{air} denotes boundaries where heat is lost to the surroundings. Similar equations exist for the backing plate but without the heat source term.

The optimization variables are denoted $\underline{\xi}$ and it is assumed that $\underline{\xi}$ is only connected to h_{bottom}. Similar considerations can be used to obtain sensitivity equations in case ξ_i is, e.g., u_w or ω.

Using a direct method the sensitivities of the temperature with respect to the optimization variables ξ_i can be written

$$\begin{aligned} 0 = &\int_\Omega \nabla w k \nabla \frac{\partial T}{\partial \xi_i} + \int_\Omega w \rho c_p \mathbf{u} \nabla \frac{\partial T}{\partial \xi_i} + \int_{\Gamma bottom} w \left[h_{bottom} \left(\frac{\partial T}{\partial \xi_i} - \frac{\partial T_2}{\partial \xi} \right) + \frac{\partial h_{bottom}}{\partial \xi_i}(T - T_2) \right] \\ &+ \int_{\Gamma air} w h_{air} \frac{\partial T}{\partial \xi_i} - \int_{\Gamma_{heat\ input}} w \left(\omega r \frac{d\tau}{dT} \frac{\partial T}{\partial \xi_i} - h_{tool} \frac{\partial T}{\partial \xi_i} \right) \end{aligned} \tag{6}$$

and for the backingplate

$$\begin{aligned} 0 = &\int_\Omega \nabla w_2 k \nabla \frac{\partial T_2}{\partial \xi_i} + \int_\Omega w_2 \rho c_p \mathbf{u_2} \nabla \frac{\partial T_2}{\partial \xi_i} + \int_{\Gamma air} w_2 h_{air} \frac{\partial T_2}{\partial \xi_i} \\ &+ \int_{\Gamma bottom} w_2 \left[h_{bottom} \left(\frac{\partial T_2}{\partial \xi_i} - \frac{\partial T}{\partial \xi_i} \right) + \frac{\partial h_{bottom}}{\partial \xi_i}(T_2 - T) \right] \end{aligned} \tag{7}$$

where w_2 is the testfunction for the backingplate. The sensitivity equations include the term $\frac{\partial h_{bottom}}{\partial \xi_i}$ which takes into account how $h_{bottom}(x, y, \underline{\xi})$ is parametrized. The above equations can be solved for $\frac{\partial T}{\partial \xi_i}$ to be used for determining gradients of temperature-based objective functions for use in gradient-based optimization techniques.

3 Optimization problems

The temperature in four points in the workpiece was recorded as function of time during a welding experiment. As the welding speed is known, the time can be translated to positions relative to the tool to be used in the steady state model at hand. The points, when translated to the Eulerian framework, in which the temperature is measured and calculated are located in four rows of seven points. The positions

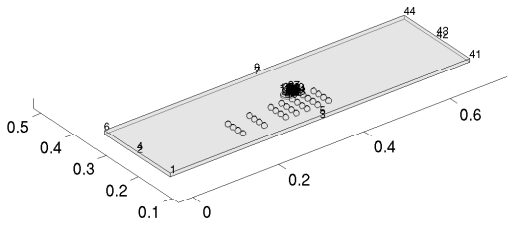

Figure 2: Points, indicated by circles, where the calculated and measured temperatures are compared. Note that more points are used on the trailing side of the tool where the effect of h_{bottom} on the temperature is largest.

are shown in Figure 2. It is recognized that the experimental data is fairly limited and this may affect the obtained results.

The measured temperatures are shown in Figure 3. The row closest to the tool clearly has artificial kinks, at $x \approx -0.03$ and $x \approx 0.05$, which are due to edge effects caused by the finite plate length in the experiments. Therefore, only three points are used from this curve for the inverse modelling, leaving $3 + 7 + 7 + 7 = 24$ points where the measured and calculated temperatures are compared.

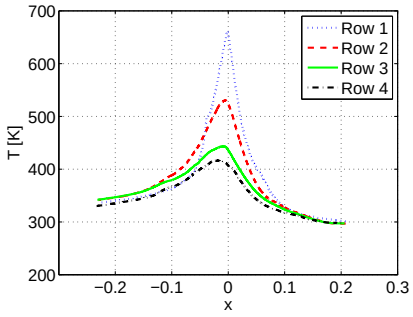

Figure 3: Experimentally measured temperatures. The row number corresponds to the number of rows of the point parallel to the welding direction in Figure 2 with row 1 closest to the tool. Note that $x = 0$ corresponds to the tool center position. The data is from [12].

One of the difficulties in using inverse modelling on the problem at hand is the parametrization of h_{bottom}. Ideally it should be allowed to take any value in any point in order for the optimization algorithm to determine the optimal spatial distribution. This could for instance be obtained by treating it as a topology optimization [6] problem in which h_{bottom} in each element on the boundary is treated as an optimization variable. However, this gives rise to some potential problems. Far from the tool, and

especially in front of the tool, the effect of the heat source is very small such that the temperature is almost the same in both the workpieces and the backingplate, meaning that the term $h_{bottom}(T - T_2)$ is very small even for large values of h_{bottom}. Thus, no matter what value is used in this area it will influence the global temperature field very little. We therefore restrict the possible distribution of h_{bottom} to some reasonable shape. As a first approximation, and in line with the literature, we use a uniform value of $h_{bottom} = h_1$. Based on the work by [13] and the fact that a large axial load is applied to the tool during welding it seems reasonable that h_{bottom} is largest under the tool and decreases with increasing distance from the tool. We therefore test a second parametrization where h_{bottom} is given by

$$\begin{aligned} h_{bottom}(x, y) &= h_1 + h_2 \exp(-h_3 r^2) \\ r^2 &= (x - x_c)^2 + (y - y_c)^2 \end{aligned} \tag{8}$$

i.e. a constant, h_1, plus a Gaussian term centered at the tool center (x_c, y_c).

Finally, we test a parametrization almost identical to the one used in [13], i.e. we divide the contact surface into six distinct zones and assume a constant h_{bottom} in each of them, see Figure 4. The optimization variables are the magnitude of h_{bottom} in each zone. The parametrization is chosen such that $h_{bottom} = h_1 + h_2 + h_6$ in zone 6 and $h_{bottom} = h_1 + h_2$ in zone 1 while the other zones are non-overlapping.

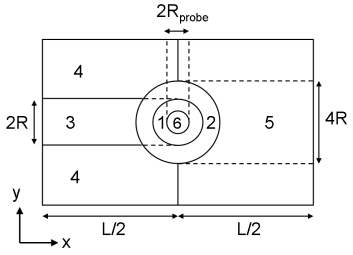

Figure 4: The division of the workpiece bottom surface into 6 zones, as indicated by the numbers. R is the shoulder radius, L the length of the model and R_{probe} the radius of the probe. Note that h_{bottom} is symmetric by construction and that the scaling is not correct.

The objective is to minimize the difference in temperature from calculations (T) and measurements (T^m). Thus we want to solve the following problem

$$\min_{\underline{\xi}} \|\underline{T}(\underline{\xi}) - \underline{T}^m\| \tag{9}$$

where $\underline{T} = \left[T_1(\underline{\xi}), ..., T_N(\underline{\xi})\right]$ and $\underline{T}^m = [T_1^m, ..., T_N^m]$ are vectors of calculated and measured temperatures, respectively, in N points.

This optimization problem may be reformulated mathematically in different ways with one possibility being

$$\begin{aligned} \min_{\underline{\xi}, \tau} \quad & \sum_{i=1}^{N} \tau_i \\ \text{s.t.} \quad & -\tau_i \leq T_i(\underline{\xi}) - T_i^m \leq \tau_i \qquad , \quad i = 1, ..., N \end{aligned} \tag{10}$$

This corresponds to minimizing the summation of the absolute temperature differences, the 1-norm, and results in a large number of constraints. However, only derivatives with respect to $\underline{\xi}$, and not τ, are expensive so the computational burden is not too great. As we consider the temperature in 24 points (and have six or less optimization variables), a direct sensitivity analysis is cheaper than an adjoint approach.

An alternative formulation is to use the squared temperature differences, i.e. a 2-norm,

$$\begin{aligned} \min_{\underline{\xi},\tau} \quad & \sum_{i=1}^{N} \tau_i \\ \text{s.t.} \quad & (T_i(\underline{\xi}) - T_i^m)^2 \leq \tau_i \qquad , \quad i = 1, ..., N \end{aligned} \tag{11}$$

where only the upper bounds on the constraints are needed. One should keep in mind that the larger exponent on the temperature difference is used, i.e. 1 in eq. (10) and 2 in eq. (11), the more influence the maximum temperature difference gets. Incorrect measurements may that way corrupt the results. Using the ∞-norm could be problematic for that reason and it has not been done here. As we are not using an equal number of points in each row of measurement points, the objective function is scaled such that equal weight is put on reducing the temperature difference in each row. All optimization problems in this article are solved using SNOPT [7].

4 Results

In this section we show results of the optimization for the three different h_{bottom} parametrizations and two different optimization problem formulations considered.

Uniform h_{bottom}

In this example h_{bottom} is considered uniform across the contact area. This is in line with the most common approach used in the literature, e.g. [12, 1]. Table 2 shows the results obtained using both formulation (10) and (11).

Table 2: Results of optimization in the case of uniform $h_{bottom}(x, y, \underline{\xi}) = h_1$. The total heat loss is the heat transfer from the workpiece to the backingplate, i.e. $\int_{\Gamma_{bottom}} h(T - T_2)dA$, and the total power is given by $\int_{\Gamma_{heat\,input}} \omega r \tau(T) dA$ where $\Gamma_{heat\,input}$ is the tool-workpiece contact area.

Opt. formulation	h_1	Obj. value	Total heat loss	Total power
Eq. (10)	415.3 $\frac{W}{m^2K}$	27.5	1523.0 W	1840.5 W
Eq. (11)	425.7 $\frac{W}{m^2K}$	295.3	1531.0 W	1847.4 W

We note that the results of the two different problems are fairly similar in terms of the optimal h_1 value. This means that also the heat loss and input power are similar while the objective function values can not be directly compared.
Figure 5 shows $\frac{\partial T}{\partial h_{bottom}}$ for the workpiece and the case of uniform $h_{bottom} = 425\,\frac{W}{m^2K}$.

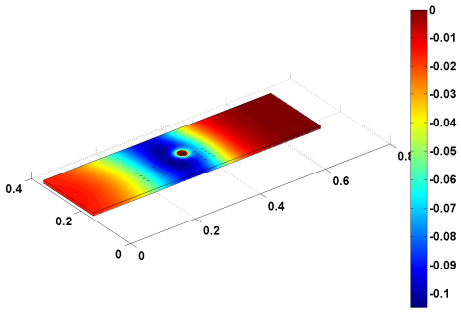

Figure 5: The figure shows $\frac{\partial T}{\partial h_{bottom}}$ for the case of uniform $h_{bottom} = 425\ \frac{W}{m^2K}$. Note that $\frac{\partial T}{\partial h_{bottom}} \approx 0$ in front of the tool, at the tool and far behind the tool, and < 0 in other areas.

It is interesting to note that $\frac{\partial T}{\partial h_{bottom}}$ is close to zero at the tool. This is caused by the (temperature dependent) TPM heat source that is self-equilibrating such that decreasing temperatures results in a larger heat input, and vice versa, and a near constant temperature field. Futhermore $\frac{\partial T}{\partial h_{bottom}}$ is very close to zero in a large region in front of the tool where the effect of the heat source is very small such that a change in h_{bottom} has very little effect on the temperatures. Also there is a region far behind the tool where h_{bottom} is small. Note that $\frac{\partial T}{\partial h_{bottom}} \leq 0$ meaning that an increase in h_{bottom} decreases the workpiece temperature as would be expected. The opposite situation holds for the backingplate such that an increase in h_{bottom} increases the backingplate temperature. It should be mentioned that Figure 5 is obtained for a specific welding setup and given h_{bottom} and as such is not general, although the conclusion that $\frac{\partial T}{\partial h_{bottom}} \leq 0$ for the workpiece, and opposite for the backingplate, is general. From the plot it might be expected that it would be best to perform temperature measurements, if the goal is to study h_{bottom}, in the region where $\frac{\partial T}{\partial h_{bottom}}$ is largest in order to have the highest sensitivity to changes.

Tool centered Gaussian

In this example we use a non-uniform h_{bottom} given by

$$\begin{aligned} h_{bottom}(x, y, \underline{\xi}) &= h_1 + h_2 \exp(-h_3 r^2) \\ r^2 &= (x - x_c)^2 + (y - y_c)^2 \end{aligned} \tag{12}$$

This parametrization means that h_{bottom} is given by a constant term plus a part of variable magnitude and range controlled by the exponential term centered at (x_c, y_c), i.e. the tool center. Results obtained using the formulations (10) and (11) are shown in table 3 and Figure 6.

Table 3: Results of optimization for $h_{bottom}(x, y, \underline{\xi}) = h_1 + h_2 \exp(-h_3 r^2)$ and problem formulation (10). Note that the constant term h_1 is fairly close to, although lower than, the result from table 2 where a uniform h_{bottom} was used.

Opt. formulation	h_1	h_2	h_3	Obj. value	Total heat loss	Total power
Eq. (10)	350.7 $\frac{W}{m^2K}$	970.7 $\frac{W}{m^2K}$	1544.4 $\frac{}{m^2}$	19.3	1673.4 W	2002.3 W
Eq. (11)	343.3 $\frac{W}{m^2K}$	962.7 $\frac{W}{m^2K}$	1354.4 $\frac{}{m^2}$	165.8	1678.1 W	2007.9 W

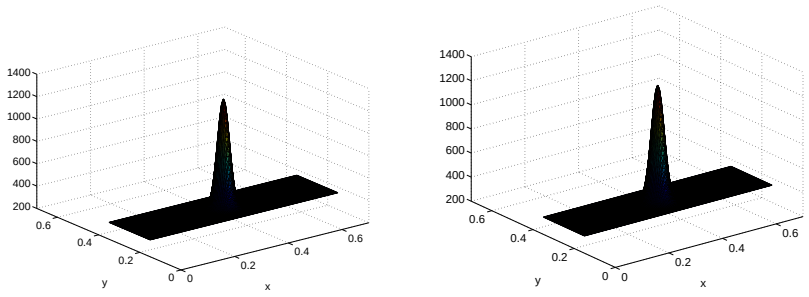

Figure 6: The optimal distribution of $h_{bottom}(x, y, \underline{\xi}) = h_1 + h_2 \exp(-h_3 r^2)$. Left: Formulation (10). Right: Formulation (11). Note that the results are almost identical.

As in the case of the uniform heat transfer coefficient the results obtained using the two different formulations are very similar in terms of the distribution of h_{bottom} as well as the resulting heat losses and input power.

When comparing the results presented so far, it is noted that for the first formulation of the optimization problem, eq. (10), the objective function value decreases from 27.5 to 19.3 when going from the uniform h_{bottom} to the non-unifom. For the second formulation the corresponding numbers are 295.3 and 165.8. These numbers correspond to decreases of 30% and 44%. This shows that much better fits between calculated and measured temperatures can be obtained by using a non-uniform h_{bottom}.

In order to study more closely the optimized h_{bottom} it is evaluated along the weldline between the plates and shown in Figure 7, for the case of formulation (10).

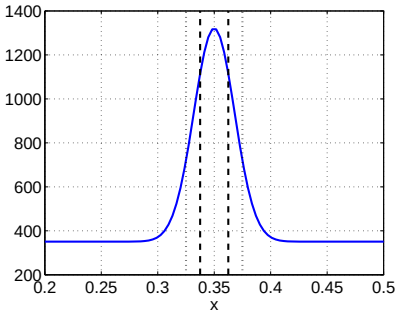

Figure 7: The optimal distribution of $h_{bottom}(x, y, \underline{\xi}) = h_1 + h_2 \exp(-h_3 r^2)$ evaluated in the weldline between the two plates. The result is obtained using formulation (10), see table 3. The dashed lines correspond to the shoulder diameter while the dotted lines correspond to 2 times the shoulder diameter. The value of h_{bottom} at the shoulder edge is $h_{bottom} \approx 1115 \frac{W}{m^2 K}$, while at a distance of 2 shoulder radii it is $h_{bottom} \approx 725 \frac{W}{m^2 K}$.

h_{bottom} divided into zones

The results presented next are obtained by introducing additional optimization variables and dividing the contact surface between workpiece and backingplate into distinct zones. Table 4 and Figures 8 and 9 show the results.

Table 4: Results of the optimization for the case where h_{bottom} is divided into 6 zones. h_1 corresponds to zone 1 in Figure 4 and so on. * means that the variable is on its lower bound. Note that in the case of formulation (10) the objective function value is higher than for the example shown in table 3 while the results obtained using formulation (11) shows a decrease in the objective function value.

Opt. form.	h_1	h_2	h_3	h_4	h_5	h_6
Eq. (10)	$1^* \frac{W}{m^2K}$	$1232.5 \frac{W}{m^2K}$	$1^* \frac{W}{m^2K}$	$415.6 \frac{W}{m^2K}$	$373.9 \frac{W}{m^2K}$	$1^* \frac{W}{m^2K}$
	Obj. value		Total heat loss		Total power	
	19.6		1687.8 W		2015.6 W	

	h_1	h_2	h_3	h_4	h_5	h_6
Eq. (11)	$5205.2 \frac{W}{m^2K}$	$360.2 \frac{W}{m^2K}$	$5399.5 \frac{W}{m^2K}$	$266.9 \frac{W}{m^2K}$	$434.4 \frac{W}{m^2K}$	$1.037 \cdot 10^5 \frac{W}{m^2K}$
	Obj. value		Total heat loss		Total power	
	137.8		2032.5 W		2361.9 W	

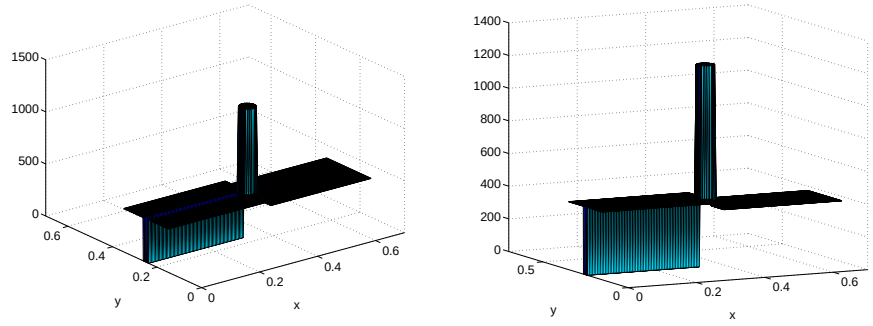

Figure 8: The optimized $h_{bottom}(x, y, \underline{\xi})$ distribution for the case where h_{bottom} is divided into 6 zones and formulation (10) is used.

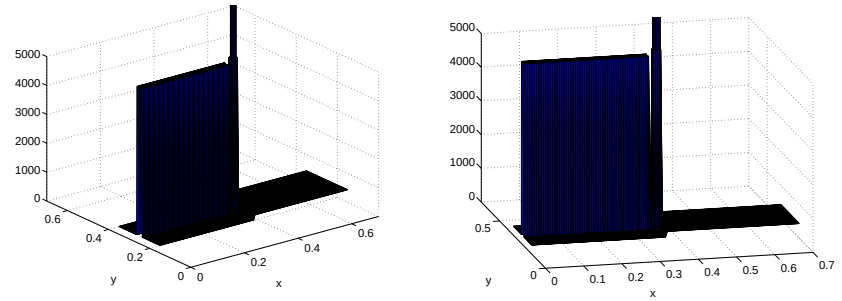

Figure 9: Optimal $h_{bottom}(x, y, \underline{\xi})$ distribution when h_{bottom} is divided into 6 zones. Formulation (10) is used. Note that in the h_{bottom} plots the z-values have been cut off at $h_{bottom} = 5000$ in order to show more clearly the spatial variation.

The results obtained using formulation (10) are as would be expected in the sense that there is a large h_{bottom} value in a region around the tool (h_2). However, behind the tool (zone 3) h_{bottom} takes its

minimum allowed value which is surprising because this area has experienced a very large pressure and relatively good contact would be expected. The minimum allowed value is set to 1 $\frac{W}{m^2K}$ in order to avoid non-physical negative values. Note that the objective function value is higher than the one obtained from the tool centered Gaussian, table 3, with only three optimization variables. Thus, the present result could be a local minimum.

These results obtained using formulation (11) are very different to those obtained using the first formulation in that h_{bottom} is very large in the region behind the tool (zone 3). Also there is an extremely large value in the probe-centered region (zone 6). However, this large h_{bottom} value acts on a very small area so most likely it does not influence the overall temperature field much, but still reduces the objective function value slightly. In this case the objective function value is reduced compared to the other two h_{bottom} parametrizations and the results look more as would be expected.

Also it is interesting to note that the total heat input and the heat loss from the workpiece increases as the number of optimization variables is increased. Part of this may be due to the use of the non-constant heat input using the TPM model, eq. (4), that responds to colder temperatures, e.g. because of higher h_{bottom} values, by increasing the heat input thereby also increasing the temperature. It would be interesting to experimentally determine the heat input, e.g. through torque measurements, as well as the temperature field in order to have more experimental data for the inverse modelling. That way both temperatures and heat input may be included in the objective function.

In Figure 10 the actual heat transfer, given by $h_{bottom}(T-T_2)$, from the workpiece to the backingplate is shown for the case of the Gaussian h_{bottom} distribution and the case of h_{bottom} divided into 6 zones based on the formulation (10). The result thus corresponds to that presented in tables 3 and 4.

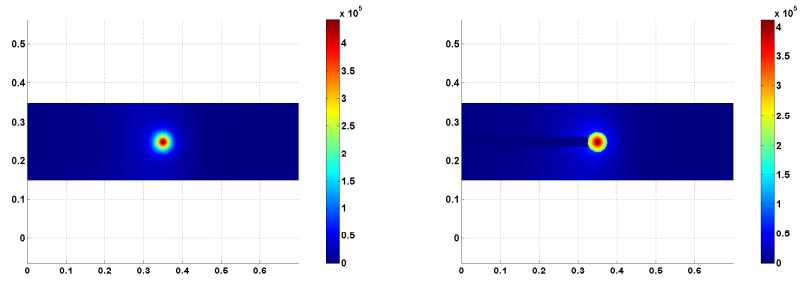

Figure 10: The figure shows $h_{bottom}(T - T_2)$, i.e. the heat flux from workpiece to backingplate. Left: Gaussian h_{bottom} distribution. Right: h_{bottom} divided into six zones.

When looking at the two figures we note that the heat flux in both cases takes its maximum value at the tool, as would be expected, and that the value is approximately the same in both cases. Comparing the heat flux for the Gaussian case to the heat transfer coefficient shown in Figure 6 we see that the distribution of the heat transfer coefficient is reflected by the heat flux. On the other hand, the more complex distribution of h_{bottom} used in the other case is not directly transferred to the heat flux. The zone behind the tool with a very low value of h_{bottom} is to some extend seen in the heat flux plot in the vicinity of the tool. The reason for the similarity is the term $T - T_2$ that is multiplied on the heat transfer coefficient to obtain the heat flux. This term is small for points far from the tool and therefore the heat flux becomes small. Thus it is found that the simpler Gaussian distribution may be adequate for modelling the non-uniform heat transfer coefficient.

The results obtained using the six variable parametrization seem to be less reliable than the results obtained using the first two parametrizations in the sense that the results are very sensitive to the choice of formulation and the fact that the objective function value in one case is larger than in the example using the Gaussian distribution. This may be a result of the limited experimental data available and as such more accurate and reliable results may be obtained if more experiments were performed. The results obtained using the Gaussian distribution on the other hand seem reliable, both because they are supported by intuition about the process, in the sense that h_{bottom} should be largest under the tool and

that they are almost identical for the two different problem formulations.

Based on the above considerations a final example is made where h_{bottom} is given by a constant plus an additional constant term close to the tool, i.e. a reduced version of the 6 parameter case studied above,

$$\begin{aligned} h_{bottom} =& h_1 + h_2 \quad \text{if } \mathrm{r} = \sqrt{(\mathrm{x} - \mathrm{x_c})^2 + (\mathrm{y} - \mathrm{y_c})^2} \leq \mathrm{R}^* \\ =& h_1 \qquad \text{else} \end{aligned} \tag{13}$$

where two different R^* values are used in the following; $R^* = R_{shoulder}$ and $R^* = 2R_{shoulder}$. This parametrization is motivated by the fact that it is easily implemented in a numerical model while still modelling the main effect, i.e. the increased h_{bottom} value under the tool. The optimized results are presented in table 5.

Table 5: Results of the optimization in the case h_{bottom} given by eq. (13).

Opt. form.	R^*	h_1	h_2	Obj. value	Total heat loss	Total power
Eq. (10)	$R_{shoulder}$	411.4 $\frac{W}{m^2K}$	4343.6$\frac{W}{m^2K}$	21.4	1801.9 W	2128.2 W
Eq. (11)	$R_{shoulder}$	427.3 $\frac{W}{m^2K}$	6775.3$\frac{W}{m^2K}$	204.4	1877.9 W	2204.4W
Eq. (10)	$2R_{shoulder}$	379.2 $\frac{W}{m^2K}$	828.0$\frac{W}{m^2K}$	20.0	1691.6 W	2017.3 W
Eq. (11)	$2R_{shoulder}$	365.6 $\frac{W}{m^2K}$	991.9$\frac{W}{m^2K}$	167.2	1711.7 W	2039.9 W

Comparing the results to those obtained using the Gaussian distribution we note that the objective function value in all cases is higher, as would be expected due to the reduced number of optimization variables and the fact that R^* is kept fixed. Furthermore, we note that the constant term is larger than the constant term of the Gaussian in the previous examples. Also the value of h_2 is significantly lower in the case of $R^* = 2R_{shoulder}$, than in the case of $R^* = R_{shoulder}$, which is caused by the fact that it has effect on a larger area. Comparing to Figure 7 the results look as wood be expected.

From an implementation point of view it is interesting to note that the objective function values in the case of $R^* = 2R_{shoulder}$ is almost as low as in the case of the Gaussian distribution, see table 3. Since the objective function value is only decreased slightly by use of more complex parametrizations it is recommended to use the two parameter distribution, i.e. such that a higher value of h_{bottom} is used in an area of radius $2R_{shoulder}$ around the heat source.

5 Conclusion

Using experimental temperature measurements the workpiece-backingplate heat transfer coefficient h_{bottom} has been studied. Four different parametrizations of h_{bottom} was tested; one uniform, one with a single peak at the tool, one where h_{bottom} was divided into 6 zones, and finally a two-parameter distribution. Most of the results were in line with expectations in the sense that h_{bottom} takes a large value in the region close to the tool where the contact pressure is large. The results were less conclusive for h_{bottom} behind the tool. We note that the tested parametrizations of h_{bottom} represent only a very limited number of the possibilities. Other parametrizations could e.g. be based on splines or radial basis functions. Also different formulations of the optimization problem could be tested.

We also note that the experimental data were fairly limited and that more data may help in obtaining better and more accurate results. However, it was found that using a non-uniform h_{bottom} it was possible to obtain much better agreement between the calculated and the measured temperatures, even for a very simple 2 variable parametrization of the heat transfer coefficient that can easily be implemented in existing models.

References

[1] Y.J. Chao and X. Qi. Thermal and thermo-mechanical modeling of friction stir welding of aluminum alloy 6061-T6. *Journal of Materials Processing & Manufacturing Science*, 7:215–233, 1998.

[2] Y.J. Chao, X. Qi, and W. Tang. Heat transfer in friction stir welding - experimental and numerical studies. *Journal of manufacturing science and engineering*, 125:138–145, 2003.

[3] P.A. Colegrove, H.R. Shercliff, and P.L. Threadgill. Modelling and development of the trivex(tm) friction stir welding tool. *4th International Symposium on Friction Stir Welding*, 2003.

[4] Comsol. www.comsol.com, 2006.

[5] T. Dickerson, Q. Shi, and H.R. Shercliff. Heat flow into friction stir welding tools. *4th International symposium on friction stir welding*, 2003.

[6] M.P. Bendsøe and O. Sigmund. *Topology Optimization: Theory, Methods and Applications*. Springer Verlag, 2003.

[7] P.E. Gill, W. Murray, and M.A. Saunders. Users guide for snopt version 7: Software for large-scale nonlinear programming, 2006.

[8] M.Z.H. Khandkar, J.A. Khan, and A.P. Reynolds. Input torque based thermal model of friction stir welding of Al-6061. *6th International trends in welding research conference proceedings, 15-19 April 2002, Pine Mountain, GA*, 2003.

[9] M.Z.H. Khandkar, J.A. Khan, A.P. Reynolds, and M.A. Sutton. Predicting residual stresses in friction stir welded metals. *Journal of materials processing technology*, 174:195–203, 2006.

[10] R. Nandan, B. Prabu, A. De, and T. Debroy. Improving reliability of heat transfer and materials flow calculations during friction stir welding of dissimilar aluminum alloys. *Welding journal*, 86:313–322, 2007.

[11] H. Schmidt and J. Hattel. A local model for the thermomechanical conditions in friction stir welding. *Modelling and simulation in materials science and engineering*, Vol. 13:77–93, 2005.

[12] H.B. Schmidt and J.H. Hattel. Thermal modelling of friction stir welding. *Scripta Materialia*, 58:332–337, 2008.

[13] V. Soundararajan, S. Zekovic, and R. Kovacevic. Thermo-mechanical model with adaptive boundary conditions fro friction stir welding of Al 6061. *International journal of machine tools & manufacture*, 45:1577–1587, 2005.

Appendix E

Maximization of advance per revolution

The following appendix deals with maximization of the advance per revolution, APR, or weld pitch given by

$$APR = \frac{u_w}{\omega}. \tag{E.1}$$

APR is a measure of how far the tool traverses during one rotation. It is desirable to maximize APR as the ratio is linked to the wear on the tool, see Tutum et al. (2009) where it is shown that the wear path is linearly proportional to APR. The model used here is a 3D thermal model consisting of the workpiece and a steel backingplate with a rotational velocity field enforced in a shear layer around the probe. This does not allow convection through the probe hole and in that sense accounts for the effect of the material flow on the thermal field, see Articles 3 and 4 for details on the velocity field.

Since we are dealing with a purely thermal model, solving an unconstrained optimization problem with the objective of maximizing APR will lead to a trivial solution where the optimum is found at the maximum allowable welding speed and minimum allowable rotational speed. Therefore, in order to secure physically reasonable results a constraint must be included. This could for example be a constraint that uses characteristics of the material flow around the tool to secure that void formation does not occur. This would require a flow model. Here, we add a constraint on the material yield strength in front of the tool, $\sigma_{y,tool}$, that should be reduced in order to allow easy traversing, and acts as a limiting factor for the welding speed. Also, a low yield stress of the material in contact with the tool will reduce the toolwear. The optimization problem to be solved is

$$\begin{aligned} \max_{u_w,\omega} \quad & APR \\ \text{s.t.} \quad & \sigma_{y,tool} \leq \tilde{\sigma} \end{aligned} \tag{E.2}$$

where $\sigma_{y,tool}$ depends on the pointwise temperature as well as the microstructural changes as defined in section 1.2.2, i.e.

$$\sigma_{y,tool}(T, X_d) = \left(1 - \frac{T - T_0}{T_m - T_0}\right)(\sigma_{y,min} + (1 - X_d)(\sigma_{y,max} - \sigma_{y,min})) \tag{E.3}$$

where the first term takes into account the pointwise temperature while the second term contains information about the thermal history, with X_d given by eq. (1.21). The cal-

$\tilde{\sigma}$	APR	u_w^*	ω^*	$1 - \frac{T-T_0}{T_m-T_0}$	$1 - X_d$
$15\,MPa$	$0.17\,mm/rev$	$3.20\,mm/s$	$1118\,RPM$	0.178	0.032
$10\,MPa$	$0.13\,mm/rev$	$5.79\,mm/s$	$2736\,RPM$	0.128	0.0

Table E.1: Results for the optimization problem (E.2) for two different constraint values $\tilde{\sigma}$. The unit for the objective function value, APR, is mm per revolution. The last two columns show the normalized temperature contribution and the microstructural softening, respectively, used in determining the yield stress in front of the tool, eq. (E.3). In both cases the yield stress constraint in front of the tool, eq. (E.2), is active and thereby the limiting factor for the maximization of APR.

culation of X_d is made easy by the fact that a point *in front of* the tool is studied, such that no considerations have to be made regarding the rotational velocity field close to the tool. However, for warm welds, e.g. small u_w and large ω, the term $(1 - X_d)$ may become negative such that $\sigma_{y,tool}(X_d) = (\sigma_{y,min} + (1 - X_d)(\sigma_{y,max} - \sigma_{y,min})) < \sigma_{y,min}$ which is not physically possible. This problem may be overcome simply by replacing $(1-X_d)$ by $\max(0, (1-X_d))$ when modelling. For the function to be differentiable, for use in gradient-based optimization, the term $(1-X_d)(\frac{1}{2} - \frac{1}{2}\tanh(K(1 - X_d))$ is used instead. For large K this term acts as a smoothed Heaviside function thereby limiting $(1 - X_d)$ to non-negative values.

The optimization results are shown in table E.1. Note that when the constraint is tightened, i.e. $\tilde{\sigma}$ is reduced from $15\,MPa$ to $10\,MPa$ the optimizer does not just reduce the welding speed u_w in order to increase the temperature. Instead the welding speed and the rotational speed is increased, but the rotational speed by a larger fraction than u_w. The optimal rotational speed in the second case is fairly large and above values normally used in FSW. The example shows how optimization methods may give results that are not readily obtained using intuition. From Table E.1 we also note that the decrease in $\sigma_{y,tool}$, in order to satisfy the constraint, has been obtained by a combination of a higher point temperature, thus reducing $1 - \frac{T-T_0}{T_m-T_0}$, and a slightly reduced fraction of hardening particles.

Appendix F

Trust-region and filter methods

This appendix presents different trust region and filter methods. Trust regions are applied to the space mapping technique in Article 1 while a filter in combination with a trust region is used in Article 3.

Trust regions are used in connection with optimization using surrogate models in order to limit the allowable step lengths. In space and manifold mapping the accuracy of the mapped coarse model, in the sense of similarity with the fine model, may be limited, especially in the first iterations. This can cause the optimization method, working on the mapped coarse model, to take large steps that do not reduce the fine model objective function value.

Trust regions are used when solving an optimization problem such as

$$\min_{\mathbf{x}} \quad \Phi(\mathbf{f}_f(\mathbf{x})) \tag{F.1}$$

through a sequence of problems using approximate models $\mathbf{f}_{c,k}$ where the subscript k indicates that $\mathbf{f}_{c,k}$ depends on the iteration number k. The approximate model could be a quadratic model when solving a general non-linear optimization problem or, in the surrogate optimization cases at hand, it could be a mapped coarse model. A constraint on the steplength $\mathbf{s} = \mathbf{x} - \mathbf{x}_k$ is enforced in the approximate optimization problem such that

$$\begin{aligned} \min_{\mathbf{x}} \quad & \Phi(\mathbf{f}_{c,k}(\mathbf{x})) \\ \text{s.t.} \quad & \|\mathbf{s}\| \leq \Delta_k^{TR} \end{aligned} \tag{F.2}$$

where Δ_k^{TR} is the (iteration number dependent) trust region size and $\|\cdot\|$ is the l_2 norm, Alexandrov et al. (1998); Walmag and Delhez (2005), or the l_∞ norm, Fletcher et al. (2002).

The size of the trust region is usually updated based on a scheme taking into account the accuracy of the approximate model by comparing the predicted improvement to the actual improvement obtained in the fine model when taking the step $\mathbf{s}$. This is done by calculating the ratio

$$\rho_k = \frac{\Phi(\mathbf{f}_f(\mathbf{x}_k)) - \Phi(\mathbf{f}_f(\mathbf{x}_{k+1}))}{\Phi(\mathbf{f}_{c,mapped}(\mathbf{x}_k)) - \Phi(\mathbf{f}_{c,mapped}(\mathbf{x}_{k+1}))} \tag{F.3}$$

Normally only succesful steps, such that $\Phi(\mathbf{f}_f(\mathbf{x}_{k+1})) < \Phi(\mathbf{f}_f(\mathbf{x}_k))$ are accepted, and based on the value of ρ_k the trust region size Δ_{k+1}^{TR} may be increased or kept fixed.

If a step is not accepted Δ_{k+1}^{TR} is decreased. In Walmag and Delhez (2005) different approaches to the update of Δ_{k+1}^{TR} are presented. A standard method is to multiply the previous trust region size with a scaling factor depending on the value of ρ_k

$$\Delta_{k+1}^{TR} = \begin{cases} \alpha_1 \Delta_k^{TR} & \text{if } \rho_k < \eta_1 \\ \Delta_k^{TR} & \text{if } \eta_1 \leq \rho_k < \eta_2 \\ \alpha_2 \Delta_k^{TR} & \text{if } \rho_k \geq \eta_2 \end{cases} \tag{F.4}$$

where the constants are chosen such that $0 < \eta_1 \leq \eta_2 < 1$ and $\alpha_1 < 1 < \alpha_2$, such that if ρ_k is close to 1 the trust region is increased by a factor α_2. Typical values are $\alpha_1 = \frac{1}{2}$ and $\alpha_2 = 2$. In Walmag and Delhez (2005) more advanced schemes using adaptive scaling factors α_i are presented.

A constrained optimization problem may be seen as a problem of minimizing the objective function value $\Phi = \Phi(\mathbf{f}_f(\mathbf{x}))$ and minimizing the constraint violation $h = h(\mathbf{f}_f(\mathbf{x}))$. The measure of constraint violation h is typically the l_1 norm of constraint violations. In trust region methods for unconstrained optimization problems, as mentioned above, only successful steps that decrease the objective function value are accepted. A filter method, Fletcher et al. (2002), may be seen as an equivalent method requiring that the step decreases either the objective function value or the constraint violation.

The filter is a list of points (Φ_i, h_i) such that no point dominates any other point in the sense that $h_i \leq h_j$ and $\Phi_i \leq \Phi_j$. A point (Φ, h) is accepted if

$$\Phi < \Phi_i \quad \text{or} \quad h < h_i \tag{F.5}$$

for all (Φ_i, h_i) belonging to the filter. If a point is accepted it is included in the filter and dominated points are removed. The filter is implemented with a trust region such that if a step results in a point $\mathbf{x}_{k+1}$ that is not acceptable to the filter, the trust region is decreased and a new $\mathbf{x}_{k+1}$ is calculated.

In practice a so-called *slanting* filter is used, in which a point is acceptable if

$$\Phi + \gamma h \leq \Phi_i \quad \text{or} \quad h \leq \beta h_i \tag{F.6}$$

with $0 < \gamma < \beta < 1$ and γ close to zero and β close to one. The slanting filter ensures that the set of unacceptable points for the updated filter includes the corresponding set for the old filter. In Fletcher et al. (2002) it is shown that the slanting filter forces iterates towards feasibility, i.e. $h = 0$, for the quadratic model considered in that article.

A filter technique is used in Article 3 in connection to a constrained manifold mapping algorithm. The motivation for this, was that the regularized version of the manifold mapping algorithm, e.q. (2.19), does not take constraints into account and that during the iterations, steps that significantly increased the constraint violation were taken. In order to overcome this problem the filter method was applied such that only "good" points, in the sense that they are acceptable to the filter, are accepted.